Détermination de sondes oligonucléotidiques pour l'exploration à haut débit de la diversité taxonomique et fonctionnelle d'environnements complexes

Nicolas Parisot

UNIVERSITE BLAISE PASCAL
CLERMONT-FERRAND II
N° D.U. 2498

UNIVERSITE D'AUVERGNE
CLERMONT-FERRAND I
ANNEE 2014

ECOLE DOCTORALE DES SCIENCES DE LA VIE, SANTE, AGRONOMIE ET ENVIRONNEMENT
N° d'ordre 643

Thèse

Présentée à l'Université Blaise Pascal pour l'obtention du grade de

DOCTEUR D'UNIVERSITE
(Spécialité : Bioinformatique et Ecologie Microbienne)

Présentée et soutenue publiquement le 17 Octobre 2014

Nicolas PARISOT

DETERMINATION DE SONDES OLIGONUCLEOTIDIQUES POUR L'EXPLORATION A HAUT-DEBIT DE LA DIVERSITE TAXONOMIQUE ET FONCTIONNELLE D'ENVIRONNEMENTS COMPLEXES

Composition du jury :

Président :	Hubert CHARLES (Pr., UMR 203 BF2I, INSA, Lyon)
Rapporteurs :	Christine GASPIN (DR INRA, Unité MIA Toulouse, Castanet-Tolosan) Guy PERRIERE (DR CNRS, UMR 5558 LBBE, Université Claude Bernard, Lyon)
Examinateurs :	Catherine ENG (Dr., DGA Maîtrise NRBC, Vert-le-Petit) Engelbert MEPHU NGUIFO (Pr., UMR 6158 LIMOS, Université Blaise Pascal, Clermont-Ferrand) Pierre PEYRET (Pr., EA 4678 CIDAM, Université d'Auvergne, Clermont-Ferrand)
Directeur :	Eric PEYRETAILLADE (Dr., EA 4678 CIDAM, Université d'Auvergne, Clermont-Ferrand)
Invitée :	Corinne BIDERRE-PETIT (CR CNRS, UMR 6023 LMGE, Université Blaise Pascal, Clermont-Ferrand)

Laboratoire
« Microorganismes : Génome et Environnement »
Unité Mixte de Recherche CNRS 6023

Equipe d'Accueil 4678 « Conception, Ingénierie et Développement de l'Aliment et du Médicament »

Détermination de sondes oligonucléotidiques pour l'exploration à haut-débit de la diversité taxonomique et fonctionnelle d'environnements complexes

Résumé :

Les microorganismes, par leurs fascinantes capacités d'adaptation liées à l'extraordinaire diversité de leurs capacités métaboliques, jouent un rôle fondamental dans tous les processus biologiques. Jusqu'à récemment, la mise en culture était l'étape préliminaire obligatoire pour réaliser l'inventaire taxonomique et fonctionnel des microorganismes au sein des environnements. Cependant ces techniques ne permettent d'isoler qu'une très faible fraction des populations microbiennes et tendent donc à être remplacées par des outils moléculaires haut-débit. Dans ce contexte, l'évolution des techniques de séquençage a laissé entrevoir de nouvelles perspectives en écologie microbienne mais l'utilisation directe de ces techniques sur des environnements complexes, constitués de plusieurs milliers d'espèces différentes, reste néanmoins encore délicate. De nouvelles stratégies de réduction ciblée de la complexité comme la capture de gènes ou les biopuces ADN représentent alors une bonne alternative notamment pour explorer les populations microbiennes même les moins abondantes.

Ces stratégies à haut-débit reposent sur la détermination de sondes combinant à la fois une forte sensibilité, une très bonne spécificité et un caractère exploratoire. Pour concevoir de telles sondes plusieurs logiciels ont été développés : PhylGrid 2.0, KASpOD et ProKSpOD. Ces outils généralistes et polyvalents sont applicables à la sélection de sondes pour tout type de gènes à partir des masses de données produites à l'heure actuelle. L'utilisation d'architectures de calculs hautement parallèles et d'algorithmes innovants basés sur les *k*-mers ont permis de contourner les limites actuelles. La qualité des sondes ainsi déterminées a pu permettre leur utilisation pour la mise au point de nouvelles approches innovantes en écologie microbienne comme le développement de deux biopuces phylogénétiques, d'une méthode de capture de gènes en solution ainsi que d'un algorithme de classification des données métagénomiques. Ces stratégies peuvent alors être employées pour diverses applications allant de la recherche fondamentale pour une meilleure compréhension des écosystèmes microbiens, au suivi de processus de bioremédiation en passant par l'identification de tous types de pathogènes (eucaryotes, procaryotes et virus).

Mots clés : *bioinformatique, métagénomique, détermination de sondes, capture de gènes, biopuces, classification*

Selection of oligonucleotide probes for high-throughput study of complex environments

Abstract:

Microorganisms play a crucial role in all biological processes related to their huge metabolic potentialities. Until recently, the cultivation was a necessary step to appraise the taxonomic and functional diversity of microorganisms within environments. These techniques however allow surveying only a small fraction of microbial populations and tend to be consequently replaced by high-throughput molecular tools. While the evolution of sequencing technologies opened the door to unprecedented opportunities in microbial ecology, massive sequencing of complex environments, with thousands of species, still remains inconceivable. To overcome this limitation, strategies were developed to reduce the sample complexity such as gene capture or DNA microarrays.

These high-throughput strategies rely on the selection of sensitive, specific and explorative probes. To design such probes several programs have been developed: PhylGrid 2.0, KASpOD and ProKSpOD. These multipurpose tools were implemented to design probes from the exponentially growing sequence datasets in microbial ecology. Using highly parallel computing architectures and innovative *k*-mers based strategies allowed overcoming major limitations in this field. The high quality probe sets were used to develop innovative strategies in microbial ecology including two phylogenetic microarrays, a gene capture approach and a taxonomic binning algorithm for metagenomic data. These approaches can be carried out for various applications including better understanding of microbial ecosystems, bioremediation monitoring or identification of pathogens (eukaryotes, prokaryotes and viruses).

Keywords: *bioinformatics, metagenomics, probe design, gene capture, DNA microarrays, binning*

Remerciements

La tradition doctorale veut qu'une thèse s'ouvre sur des remerciements mais tous les mots que je trouverai au cours des prochains paragraphes ne suffiront pas à exprimer toute ma gratitude.

Je souhaite avant tout exprimer mes remerciements aux différents directeurs d'unités, Christian Amblard et Télesphore Sime-Ngando de l'UMR 6023 LMGE, et Monique Alric de l'EA 4678 CIDAM. Je vous remercie sincèrement pour votre accueil et pour m'avoir donné les moyens de réaliser cette thèse dans de bonnes conditions.

Mes remerciements les plus sincères vont ensuite à Christine Gaspin et Guy Perrière pour m'avoir fait l'honneur d'accepter d'être rapporteurs de ce travail ainsi qu'à Hubert Charles, Catherine Eng et Engelbert Mephu Nguifo pour avoir accepté d'examiner mon travail et faire partie de mon jury de thèse. Je voudrais également remercier le CNRS et la Direction Générale de l'Armement (DGA) pour la bourse de thèse et la confiance qu'ils m'ont accordée. A ce titre je remercie également Gilles Vergnaud et Emmanuelle Guillot-Combe d'avoir été tuteurs de cette thèse.

Merci également à Jérôme Salse, Lionel Ranjard, Sébastien Rimour et Philippe Leroy d'avoir accepté de participer aux différents comités de suivi de thèse. Je vous remercie pour les conseils et différentes orientations que vous avez donné à mon travail.

Mes remerciements les plus sincères et les plus chaleureux vont à Eric Peyretaillade, mon directeur de thèse pendant ces trois années mais pas seulement. Je voudrais te remercier pour tout ce que tu as fait pour moi depuis maintenant plus de six ans. Depuis mon stage de DUT et ce projet un peu farfelu de MicroAnnot, tu m'as fait confiance et m'as guidé dans mes travaux et mes choix professionnels. Ta disponibilité, ta passion et ton efficacité ont été pour moi un exemple et ont grandement contribué à l'aboutissement ce travail de thèse. Au delà de l'aspect professionnel, c'est une sincère amitié et complicité qui s'est créée et je ne te remercierai jamais assez pour nos nombreuses discussions, les pieds dans l'eau près de La Bourboule, en voiture lors des aller-retours à Aurillac ou dans un voyage un peu fou à Caen. Merci de m'avoir tant donné...

Puisque l'un ne va rarement sans l'autre, j'adresse également un grand merci à Pierre Peyret. Merci à vous pour vos précieux conseils, votre soutien infaillible et pour le goût de la recherche que vous transmettez à tous les étudiants qui passent dans cette équipe. Même si je partais avec l'handicap d'avoir été sur les bancs d'une école d'ingénieur, vous avez su me laisser une grande liberté et vous m'avez accordé les moyens matériels de mener à bien mes travaux. Un grand merci pour votre confiance qui m'a permis d'être associé aux nombreux projets de recherche de l'équipe. Merci pour tout ce que vous avez pu faire pour moi.

Comme l'heure est aux duos, je voudrais remercier Delphine Boucher et Corinne Biderre-Petit d'avoir supporté notre trio masculin. Merci pour votre aide, vos conseils, votre gentillesse et pour tous les bons moments passés ensemble et en compagnie de vos familles.

Eric Dugat-Bony, en voilà un qui a également beaucoup contribué au bon déroulement de cette thèse ! Tu as débuté ta thèse lors de mon stage de DUT et tu l'as soutenue durant mes premiers mois de thèse. Nos nombreuses discussions m'ont permis d'avancer et pendant ces trois années de thèse je n'ai eu de cesse de me rappeler ces quelques mots lâchés à la fin de ta soutenance : « Tu n'as plus qu'à faire mieux ! ». J'espère ne pas t'avoir déçu ! Un grand merci pour les agréables moments passés ensemble et je te souhaite tout le bonheur que tu mérites.

J'ai trouvé un autre ami durant cette thèse, il s'agit de ce cher Jérémie Denonfoux. Deux ans passés ensemble dans le même bureau, un soutien sans faille et une bonne humeur à toute épreuve ! Dire qu'on a fait les 400 coups avec toi et Eric serait probablement exagéré mais on en a vécu des bons moments ensemble ! Des concours de pétanque, des fléchettes à travers le bureau, des chaises de jardin en miettes, une passion pour la musique des années 80 et 90, et des répliques de films cultes ! Alors que ce soit Moundir, Thierry Pastor, Dr. Alban ou Retour vers le futur, ils seront tous synonymes de « Loulou ». Merci pour tout et je te souhaite plein de bonnes choses.

Après avoir débuté ma thèse avec ces deux acolytes masculins, je l'ai conclu avec plus de délicatesse en compagnie de Céline Ribière et Cyrielle Gasc. Merci à toutes les deux de m'avoir supporté en cette fin de thèse, merci d'avoir été là pour moi et pour nos nombreuses discussions.

La suite de mes remerciements va aux anciens membres de l'équipe G2IM du LMGE pour leur accueil et leur bonne humeur quotidienne. Un grand merci à Brigitte Chebance pour

toute son aide et sa profonde gentillesse. Merci également à Anne Moné et Isabelle Pinto pour leur sympathie. Merci à Olivier Gonçalves pour son humour, aux anciens doctorants et post-doctorants pour leurs conseils : Cécile Militon, Sébastien Terrat, Jérôme Brunellière, Mohieddine Missaoui, Emilie Dumas, Ourdia Bouzid, Sophie Comtet et Faouzi Jaziri.

Merci également à mes collègues et amis du LMGE. Tout d'abord Mylène Hugoni qui m'a vu débarquer en Master dans son immeuble de la rue Nelaton. Merci pour ton amitié, nos discussions sincères et tous les moments que nous avons passé ensemble. Notre entraide a beaucoup compté ces dernières années et je ne parle pas que de mon four ! Tu trouveras le bonheur que tu mérites. Benjamin Misson je souhaite également te remercier pour ton amitié et te féliciter pour la magnifique famille que tu as su fonder. Je ne veux pas oublier de remercier non plus Frédéric Delbac, Hicham El Alaoui, Didier Debroas et toute l'équipe MEB, Emilie Duffaut, Stéphanie Palesse, Julie Aufauvre, Mathieu Roudel, Cyril Vidau, Marion Sabart et beaucoup d'autres.

Je voudrais ensuite exprimer ma gratitude aux membres de l'EA CIDAM que j'ai pu côtoyer pour cette moitié de thèse. Je veux tout d'abord remercier Lucie Etienne-Mesmin pour sa gentillesse et son amitié. Même si nous n'avons pas pu passer beaucoup de temps ensemble j'ai énormément apprécié ta compagnie. Merci à Jonathan Thévenot, ton successeur, pour nos discussions. Je remercie aussi William Tottey pour ses passages dans le bureau et pour le souvenir d'une soirée plutôt chargée. Mes remerciements vont également à Jean-François Brugère et Guillaume Borrel pour m'avoir associé à leurs travaux et pour nos échanges enrichissants. Un grand merci à Eléonore Attard et Réjane Beugnot, et enfin merci à tous les membres de l'équipe : Charlotte Cordonnier, Mickaël Fleury, Olivier Le Goff, Aurélie Guerra, Hassana Hsein, Nadia Gaci, Emmanuelle Lainé, Xie Xiaoyu, Thierry Allario, Suenia De Paiva Lacerda, Stéphanie Blanquet-Diot, Eric Beyssac, Ghislain Garrait, Valérie Hoffart, Jean-Michel Cardot, Pascale Gauthier, Jean-François « Monsieur » Jarrige, Pierre-Charles Romond, Sylvain Denis, Sandrine Chalancon, Carine Mazal, Marie Cousseau, Stéphane Jardin, Christelle Guyard et Manon Martinet.

Je n'oublie pas non plus les nombreux stagiaires passés dans l'équipe que j'ai eu le plaisir de rencontrer ou même d'encadrer durant cette thèse : Kévin Gravouil, Pierre Marijon, Audrey Serra, Aimeric Bruno, Gaëtan Guillaume, Anne-Sophie Yvroud, Stéphane Freitas, Laura Dumas, Mélanie Mitchell, Amandine Ollivier, Maxime Ossedat, Sylvain Laperche, Sarah Orlhac, Lauriane Roux, Nicolas Gallois, Stella Baret, Emilie Girard, Valérie George,

Thomas Douëllou, Clémence Defois, Auriane Bernard, Camille Forest, Marine Bertoni, Mélanie Blanc, Fanny Matner, Mickaël Mege et enfin Florent Murat pour les bons moments passés ensemble.

Mes remerciements vont également à d'autres personnes des laboratoires clermontois avec qui j'ai partagé de bons moments : Priscilla Branchu, Bruno Lamas, Benoit Chassaing, Jennifer Raisch, Amélie De Vallée, Laureen Crouzet et bien d'autres.

J'adresse toute ma gratitude à Valérie Polonais et Abdel Belkorchia pour avoir rendu mes voyages à Aurillac encore plus agréables et pour leur soutien. J'en profite aussi pour remercier les membres de l'IUT d'Aurillac pour leur accueil.

Ces travaux de thèse ont donné lieu à de nombreuses collaborations et je souhaiterais remercier l'ensemble des personnes qui m'ont fait confiance : Diego Morgavi, Milka Popova, Pascale Mosoni, Evelyne Forano, Roland Marmeisse, Patricia Luis, Laurence Fraissinet, Claudia Bragalini, Antoine Mahul, Sébastien Cipière, Sylvain Charlat, Yannick Bidet, Robert Duran, Aurélie Cébron, Jean-Yves Richard, Paul O'Toole, Sylviane Derzelle, Emmanuelle Lerat, Nicolas Corradi, Jean-François Pombert, Chantal Vaury, Emilie Brasset, Silke Jensen, Mathilde Bonnet, Matthias Zytnicki et beaucoup d'autres.

Ensuite, mes remerciements vont à ma famille et mes amis pour leur compréhension et leur soutien. Merci à mes deux grands-mères qui malgré l'éloignement s'inquiètent des travaux de leur petit-fils.

Pour finir je voudrais te remercier d'être là pour moi et pour l'amour sincère que tu me portes. J'espère profondément pouvoir construire un avenir avec toi et si j'écris seulement ces quelques lignes pour toi c'est que tous les mots ne suffiraient pas à te témoigner mon amour et comme tu le dis si bien : « L'important c'est ce que ça représente. ».

« La valeur d'un homme tient dans sa capacité à donner et non dans sa capacité à recevoir. »

Albert Einstein

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

%GC	Pourcentage en bases Guanine et Cytosine
16S	16 Svedberg
18S	18 Svedberg
23S	23 Svedberg
25S	25 Svedberg
28S	28 Svedberg
ADN	Acide DésoxyriboNucléique
ADNc	ADN Complémentaire
ADNg	ADN Génomique
ADNr	ADN Ribosomique
ANR	Agence Nationale de la Recherche
ARDRA	*Amplified Ribosomal DNA Restriction Analysis*
A-RISA	*Automated-Ribosomal Intergenic Spacer Analysis*
ARN	Acide RiboNucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ATP	Adénosine TriPhosphate
BLAST	*Basic Local Alignment Search Tool*
CARD-FISH	*CAtalysed Reporter Deposition-FISH*
CDD	*Charge Coupled Device*
CD-HIT	*Cluster Database at High Identity with Tolerance*
CDS	*Coding DNA Sequence*
CNRS	Centre National de la Recherche Scientifique
COG	*Clusters of Orthologous Groups of proteins*
CPU	*Central Processing Unit*
CRP	*Cystein-Rich Protein*
CRRI	Centre Régional des Ressources Informatiques
CRT	*Cyclic Reversible Termination*
CSS	*Cascading Style Sheets*
CSV	*Comma-Separated Values*
Cy3	Cyanine 3
Cy5	Cyanine 5
DDBJ	*DNA Data Bank of Japan*
ddNTP	DidésoxyriboNucléotide TriPhosphate
DGGE	*Denaturing Gradient Gel Electrophoresis*
dNTP	DésoxyriboNucléotides TriPhosphate
DYP	*Dye-decolorizing Peroxydase*
EBI	*European Bioinformatics Institute*
EC2CO	Ecosphère Continentale et Côtière
EGI	*European Grid Infrastructure*
EMBL	*European Molecular Biology Laboratory*
emPCR	PCR en émulsion
ENA	*European Nucleotide Archive*
ENV	*Environmental division*
FGA	*Functional Gene Array*
FISH	*Fluorescence In Situ Hybridization*
fmdC	Sous-unité C de la formyl-méthanofurane déshydrogénase
FUN	*Fungal division*
Gb	Gigabase
GH	Glycoside Hydrolase
GHz	GigaHertz
Go	Gigaoctet
GPGPU	*General-purpose Processing on GPU*
GPU	*Graphics Processing Units*
gyrB	Sous-unité B de l'ADN gyrase
HiSpOD	*High-Specific Oligonucleotide Design*
HITChip	*Human Intestinal Tract Chip*
HOMIM	*Human Oral Microbe Identification Microarray*
HPC	*High Performance Computing*
Hsp60	*Heat Shock Protein* de 60 kDa
HTC	*High Throughput Computing*
HTML	*HyperText Markup Language*
HuGChip	*Human Gut Chip*
INRA	Institut National de la Recherche Agronomique
IUB	*International Union of Biochemistry*
IUPAC	*International Union of Pure and Applied Chemistry*
JSON	*JavaScript Object Notation*
KASpOD	*K-mer based Algorithm for high-Specific Oligonucleotide Design*
kDa	Kilo Dalton
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
kpb	Kilo paire de bases
LBBE	Laboratoire de Biométrie et Biologie Evolutive
LMGE	Laboratoire Microorganismes : Génome et Environnement
MAR-FISH	*MicroAutoRadiography-FISH*
Mb	Megabase
MCR	Méthyl-Coenzyme M Réductase
McrA	Sous-unité A de la Méthyl-Coenzyme M Réductase
McrB	Sous-unité B de la Méthyl-Coenzyme M Réductase
McrC	Sous-unité C de la Méthyl-Coenzyme M Réductase
McrD	Sous-unité D de la Méthyl-Coenzyme M Réductase
McrG	Sous-unité G de la Méthyl-Coenzyme M Réductase
MDA	*Multiple Displacement Amplification*
MLSA	*MultiLocus Sequence Analysis*
MMO	Méthane Monooxygénase
Mpb	Mega paire de bases
MPI	*Message Passing Interface*
MRT	Isoforme II de la Méthyl-Coenzyme M Réductase
MspA	Porine A de *Mycobacterium smegmatis*
NanoSIMS	Spectromètre de masse à ionisation secondaire à l'échelle nanométrique
NCBI	*National Center for Biotechnology Information*
ng	Nanogramme
NGS	*Next Generation Sequencing*
OPD	*Oligonucleotide Probe Database*
ORF	*Open Reading Frame*
OTU	*Operational Taxonomic Unit*
PatMaN	*Pattern Matching in Nucleotide database*
pb	Paire de bases
PCBs	PolyChloroByphényls

PCR	*Polymerase Chain Reaction*
Pfam	*Protein families*
pH	Potentiel d'Hydrogène
PHP	*Hypertext Preprocessor*
PhylOPDb	*Phylogenetic Oligonucleotide Probe Database*
pMMO	Forme Particulaire de la MMO
pmoA	Sous-unité A de la pMMO
POA	*Phylogenetic Oligonucleotide Array*
PPi	Pyrophosphate inorganique
PRO	*Prokaryote division*
ProKSpOD	*Protein coding sequence based K-mer algorithm for high-Specific Oligonucleotide Design*
PTP	*PicoTiterPlate*
qPCR	PCR quantitative
RAM	*Random Access Memory*
RDP	*Ribosomal Database Project*
recA	Sous-unité A de la protéine de recombinaison bactérienne
SIP	*Stable Isotope Probing*
SMP	*Symmetric MultiProcessing*
SMRT	*Single Molecule Real Time Technology*
SQL	*Structured Query Language*
SRA	*Short Read Archive*
SSCP	*Single Strand Conformation Polymorphism*
STAP	*Small subunit ribosomal RNA Taxonomy and Alignment Pipeline*
Taq	*Thermus aquaticus*
TGGE	*Temperature Gradient Gel Electrophoresis*
T_m	*Melting temperature*
To	Teraoctet
T-RFLP	*Terminal-Restriction Fragment Length Polymorphism*
UMR	Unité Mixte de Recherche
USB	*Universal Serial Bus*
WGA	*Whole Genome Array*

Introduction générale

De par la grande diversité de leurs métabolismes et leur incroyable capacité d'adaptation, les microorganismes sont retrouvés dans tous les écosystèmes même les plus extrêmes, et interviennent dans tous les processus globaux. L'écologue microbien est ainsi amené à faire l'inventaire taxonomique et fonctionnel des microorganismes afin d'évaluer la structure et la fonction des communautés microbiennes.

Cependant, l'acquisition de ces données représente un défi majeur du fait de l'extraordinaire diversité et complexité (*i.e.* structurale et fonctionnelle) des communautés microbiennes présentes au sein des différents écosystèmes qui restent encore aujourd'hui largement méconnues (Torsvik *et al.* 1990 ; Whitman *et al.* 1998 ; Hugenholtz *et al.* 1998). Jusqu'à récemment, la mise en culture était une étape obligatoire dans l'identification des microorganismes et la caractérisation de leurs capacités métaboliques. Or, ces techniques de culture ne permettent d'isoler qu'une très faible fraction des populations microbiennes. En effet, on estime que moins de 1% des microorganismes sont aujourd'hui cultivés (Amann *et al.* 1995 ; Hugenholtz *et al.* 1998 ; Rappé & Giovannoni 2003). C'est grâce à l'émergence des outils moléculaires dits à haut-débit que la caractérisation structurale et fonctionnelle des communautés microbiennes a pu être facilitée. Ainsi, l'essor de la métagénomique, lié à l'évolution du séquençage massif, a laissé entrevoir de nouvelles perspectives. Cependant, l'utilisation directe des nouvelles approches de séquençage sur des environnements complexes reste encore délicate du fait des difficultés d'interprétation des masses de données générées et des coûts restant élevés. La réduction ciblée de la complexité semble alors être une bonne alternative notamment pour explorer les populations peu abondantes. Avec l'objectif de réduire cette complexité, l'enrichissement préalable des gènes ou génomes d'intérêt représente donc une approche innovante. Parallèlement, pour appréhender les communautés microbiennes des environnements complexes, les biopuces ADN (phylogénétiques ou fonctionnelles) apparaissent également être des outils de choix du fait de leur simplicité d'utilisation, de leur capacité de multiplexage assurant la gestion simultanée d'un grand nombre d'échantillons et de la facilité d'interprétation des résultats.

Le point clé du développement de ces deux types d'approches porte sur la détermination *in silico* de sondes de haute qualité à la fois spécifiques et sensibles. L'objectif de ces travaux de thèse repose donc sur la conception de nouveaux algorithmes de détermination de sondes

oligonucléotidiques pour l'exploration taxonomique et fonctionnelle d'environnements complexes. Ces sondes, véritables signatures spécifiques de fragments d'ADN d'intérêt, devront être utilisables pour plusieurs types d'approches, aussi bien moléculaires qu'informatiques, allant des biopuces ADN à la capture de gènes en passant par des outils bioinformatiques d'annotation des données métagénomiques.

Ainsi, le mémoire de thèse sera structuré en trois parties, dont la première fera état des connaissances bibliographiques en génomique environnementale avec notamment une présentation détaillée des outils haut-débit récemment développés pour caractériser la diversité taxonomique et fonctionnelle au sein des environnements. Cette première partie se terminera par la description des différentes approches de détermination de sondes oligonucléotidiques, ces dernières étant la pierre de voûte de certaines approches moléculaires mais peuvent aussi représenter des signatures pour l'affiliation *in silico* des données produites par les séquenceurs haut-débit.

La seconde partie portera ainsi sur la présentation de nouvelles stratégies de sélection de sondes oligonucléotidiques permettant l'évaluation de la diversité microbienne connue ou encore inconnue. La troisième partie présentera, quant à elle, les diverses applications moléculaires et bioinformatiques des sondes oligonucléotidiques précédemment déterminées.

Une conclusion générale fera le bilan des avancées apportées par ces travaux de thèse et des perspectives pour l'étude des formidables capacités d'adaptation des microorganismes.

PARTIE 1 : Synthèse bibliographique

1. Génomique environnementale

1.1 Diversité microbienne et méthodes d'études

1.1.1 Biodiversité des microorganismes

La Terre abrite entre 4.10^{30} et 6.10^{30} cellules procaryotes (Whitman *et al.* 1998) qui constituent la majeure partie de la biomasse. En admettant une taille moyenne de 3 Mpb pour les génomes procaryotes, le matériel génétique de ces cellules mis bout à bout représenterait alors une distance de 100 milliards d'années-lumière soit plus que la taille de l'univers. Les microorganismes sont donc omniprésents et ce quel que soit le type d'environnement considéré. Par exemple, même chez l'être humain, les cellules bactériennes sont majoritaires. En effet, on estime que l'Homme est constitué d'environ 10^{13} cellules, mais qu'il abrite 10^{14} bactéries (Savage 1977 ; Berg 1996).

En plus de la biomasse élevée qu'ils représentent, les microorganismes (bactéries, archées et eucaryotes unicellulaires) montrent une diversité impressionnante puisque certains auteurs estiment le nombre d'espèces bactériennes à plus d'une dizaine de millions (Allsopp *et al.* 1995 ; Eisen 2007). Ils sont aussi capables de s'adapter à tous les types de milieux, y compris les plus extrêmes. Ainsi, même avec des températures comprises entre -5°C et -70°C et une teneur en eau inférieure à 2%, un gramme de sol antarctique renferme entre 10^5 et 10^9 cellules (Cowan *et al.* 2002). Par opposition, on retrouve également des espèces hyperthermophiles près de sources hydrothermales à plus de 300°C comme les fumeurs noirs (*e.g. Pyrolobus fumarii,* isolé d'une cheminée hydrothermale, est capable de se développer à 113°C (Blochl *et al.* 1997)).

Par leur abondance, leur pouvoir d'adaptation et leur grande diversité métabolique, les microorganismes jouent un rôle majeur dans l'organisation, le fonctionnement et l'évolution des écosystèmes. Ils sont à la fois producteurs, consommateurs et décomposeurs, et en intervenant dans les différentes étapes de la transformation de la matière organique, ils sont seuls capables d'effectuer certains processus de transformation. De par la multiplicité de ces potentialités métaboliques, certaines caractéristiques enzymatiques sont d'un grand intérêt pour l'Homme. C'est ainsi le cas pour celles permettant, soit de synthétiser des molécules à haute valeur ajoutée (*e.g.* antibiotiques), soit de dégrader des molécules nocives pour l'être

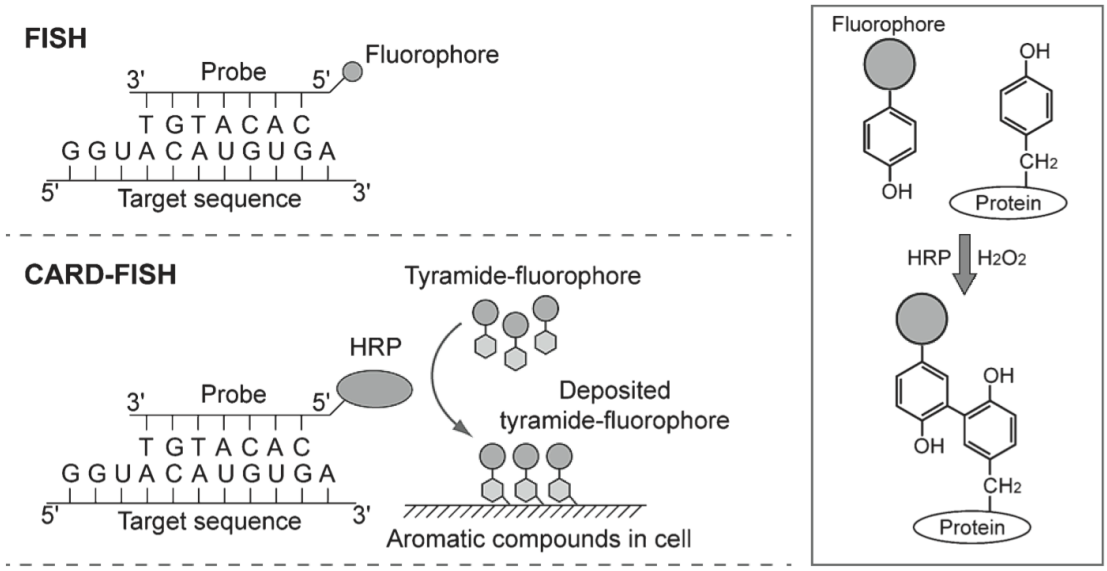


Figure 1. Représentation schématique des techniques FISH et CARD-FISH.

La technique FISH repose sur l'hybridation *in situ* de sondes marquées par un fluorophore sur des cibles d'acides nucléiques. Le CARD-FISH améliore cette technique grâce à l'amplification du signal par le tyramide. L'encadré représente la réaction d'immobilisation du tyramide sur la tyrosine. HRP : « *horseradish peroxydase* » ou péroxydase de raifort. D'après Kubota (2013).

humain (*e.g.* bioremédiation). Cependant ces potentialités sont encore largement inexplorées et inexploitées.

1.1.2 Méthodes d'étude de la diversité microbienne

L'exploration des environnements complexes demeure actuellement l'un des défis majeurs en écologie microbienne du fait de l'importante diversité des microorganismes qu'ils hébergent. Pour une bonne compréhension du fonctionnement des écosystèmes il est important i) d'identifier les microorganismes (*i.e.* structure des communautés), ii) de caractériser leurs fonctions métaboliques et enfin iii) de relier la structure à la fonction. Un grand nombre de méthodes culturales, moléculaires et biochimiques ont été dès lors appliquées pour répondre à ces objectifs.

1.1.2.a Approches basées sur la culture et l'observation cellulaire

Les méthodes classiques de microbiologie basées sur la culture impliquent l'inoculation d'un échantillon environnemental sur des milieux de culture (solides ou liquides) dont la composition doit favoriser l'isolement des microorganismes d'intérêt (Hugenholtz 2002). Les paramètres de croissance tels que le substrat, la température, le pH, le temps d'incubation, l'aération, la présence ou l'absence de lumière diffèrent selon les populations à caractériser. Les approches culturales sont mises en œuvre soit pour dénombrer les cellules cultivables et viables (Sait *et al.* 2002), soit pour sélectionner des microorganismes présentant un caractère particulier.

De manière à obtenir des données complémentaires aux approches culturales certaines méthodes d'observation ont été mises au point afin de visualiser directement l'abondance, la répartition et les interactions des communautés d'intérêt, et ce de manière *in situ*. Ces approches sont basées sur l'utilisation de sondes oligonucléotidiques fluorescentes ciblant spécifiquement des séquences d'acides nucléiques particulières. Les plus utilisées actuellement sont le *Fluorescent In Situ Hybridization* (FISH) et le CARD-FISH (DeLong *et al.* 1989 ; Amann *et al.* 1990 ; Schönhuber *et al.* 1997 ; Bottari *et al.* 2006 ; Valm *et al.* 2011 ; 2012 ; Kubota 2013) (**Figure 1**). Ces techniques génèrent des données quantitatives intéressantes mais renseignent difficilement sur les capacités métaboliques des microorganismes. Ainsi d'autres stratégies, couplant le FISH à des techniques isotopiques (MAR-FISH, FISH-NanoSIMS) (Lee *et al.* 1999 ; Li *et al.* 2008), ou encore la SIP (*Stable Isotope Probing*) incorporant des isotopes stables issus de substrats marqués au niveau des molécules d'ADN (Radajewski *et al.* 2000), ont été développées de manière à pouvoir relier

l'identification des microorganismes à leurs fonctions métaboliques (Wagner *et al.* 2006). Cependant, ces méthodes restent limitées quant à leur application du fait de leur difficulté de mise en œuvre mais surtout par leur faible débit.

1.1.2.b Méthodes basées sur l'analyse des molécules d'acides nucléiques

Le développement des techniques moléculaires ces 25 dernières années permet aujourd'hui d'aborder les problématiques d'écologie microbienne simplement au travers de l'analyse des molécules d'acides nucléiques (ADN et/ou ARN) en contournant les limites des méthodes culturales ou d'observation au microscope des microorganismes (Amann *et al.* 1995 ; Pace 1997). Ces approches indépendantes de la culture reposent sur l'utilisation de génomes entiers ou de biomarqueurs capables de renseigner sur l'identité (*i.e.* biomarqueurs phylogénétiques) ou le rôle fonctionnel (*i.e.* biomarqueurs fonctionnels) d'un grand nombre de microorganismes.

i. Utilisation des biomarqueurs

Le biomarqueur phylogénétique le plus utilisé en écologie microbienne est le gène codant pour la petite sous-unité de l'ARN ribosomique (ARNr 16S chez les procaryotes et ARNr 18S chez les eucaryotes) (Woese *et al.* 1990). Le biomarqueur phylogénétique ADNr 16S (Woese 1987) est largement utilisé en écologie microbienne pour la description des communautés bactériennes et archées de l'environnement puisque i) il est ubiquiste c'est-à-dire retrouvé chez tous les procaryotes grâce à son rôle clé dans la traduction de l'ARNm en protéine, ii) il possède une structure en mosaïque incluant des régions conservées (*i.e.* permettant son isolement) mais aussi variables et hypervariables (*i.e.* à la base des comparaisons phylogénétiques), iii) il ne subit pas ou peu de transfert horizontal et de recombinaisons (Hugenholtz 2002). De plus, sa taille adaptée (~1500 pb) ainsi que le nombre croissant de séquences codantes pour l'ARNr 16S présentes dans les bases de données, font de lui un marqueur de choix. L'analyse de ce biomarqueur peut également assurer l'identification de groupes fonctionnels comme par exemple les microorganismes déhalorespirants ou encore les sulfato-réducteurs. De la même manière, des biomarqueurs fonctionnels peuvent renseigner sur l'identité des microorganismes comme par exemple le gène *mcrA* codant pour la sous-unité α de la méthyl coenzyme M réductase chez les archées méthanogènes (Narihiro & Sekiguchi 2011) ou encore le gène *pmoA* codant pour la méthane monooxygénase et retrouvé uniquement chez les bactéries méthanotrophes (Luke & Frenzel 2011). Toutefois, les biomarqueurs fonctionnels sont généralement des gènes codant pour des

enzymes impliquées dans des métabolismes d'intérêt. D'autres biomarqueurs généralistes tels que les gènes codant pour la sous-unité β de l'ARN polymérase (*rpoB*), la sous-unité β de l'ADN gyrase (*gyrB*), la recombinase A (*recA*) ou encore la « *heat shock protein 60* » (Hsp60), ont été utilisés en écologie microbienne pour l'étude des communautés microbiennes ou encore pour différencier certaines espèces bactériennes (Santos & Ochman 2004 ; Ciccarelli *et al.* 2006 ; Case *et al.* 2007 ; Mering *et al.* 2007 ; Ghebremedhin *et al.* 2008 ; Liu *et al.* 2012b).

ii. Analyse partielle des communautés microbiennes basée sur l'amplification PCR

De nombreuses méthodes moléculaires d'analyse des communautés microbiennes utilisent la technique de réaction de polymérisation en chaîne (PCR) (Saiki *et al.* 1988) pour amplifier une région d'ADN cible grâce à un couple d'amorces. Cette méthode a révolutionné l'étude des communautés microbiennes présentes dans les environnements complexes en étant capable de cibler n'importe quelle population ou groupe de microorganismes pour lesquels des informations de séquences sont disponibles.

L'utilisation de la PCR couplée au clonage/séquençage permettent d'obtenir des données de séquences qui seront ultérieurement comparées à des bases de données généralistes comme Genbank (Benson *et al.* 2014) ou plus spécifiques des séquences d'ADNr comme SILVA (Quast *et al.* 2013), RDP (*Ribosomal Database Project*) (Cole *et al.* 2013) ou encore Greengenes (McDonald *et al.* 2012). Ainsi, les séquences de marqueurs phylogénétiques sont affiliées à différents rangs taxonomiques (allant du phylum jusqu'à l'espèce) en prenant en compte différents seuils de similarité nucléique. Bien que les banques de clones construites à partir de séquences d'ADNr 16S permettent d'explorer la diversité et d'identifier de nouveaux taxa bactériens et archéens, des études ont montré que des échantillons environnementaux, tels que les sols, requièrent plus de 40 000 clones pour décrire uniquement 50% de la diversité totale (Dunbar *et al.* 2002). Cependant, les banques de clones d'ADNr 16S construites pour les études environnementales montrent moins de 1000 séquences, et proposent donc une vision très réduite de la diversité bactérienne présente dans un échantillon.

D'autres méthodes basées sur l'amplification PCR, comme les techniques d'empreintes génétiques, donnent un profil des communautés microbiennes basé sur l'analyse directe des amplicons obtenus à partir d'ADN environnemental (Ramette 2009). Au cours de

ces 25 dernières années, un large éventail de techniques a été développé pour la description des communautés microbiennes en produisant des empreintes moléculaires basées sur des polymorphismes de séquences ou de longueurs des gènes biomarqueurs (Kirk *et al.* 2004). Parmi ces différentes techniques, il est possible de citer la *Denaturing Gradient Gel Electrophoresis / Temperature Gradient Gel Electrophoresis* (DGGE / TGGE) (Gelsomino *et al.* 1999), la *Temporal Temperature Gel Electrophoresis* (TTGE) (Muyzer *et al.* 1993), la SSCP (*Single Strand Conformation Polymorphism*) (Lee *et al.* 1996), la T-RFLP (*Terminal-Restriction Length Polymorphism*) (Liu *et al.* 1997), la ARDRA (*Amplified Ribosomal DNA Restriction Analysis*) (Liu *et al.* 1997) ou encore la A-RISA (*Automated-Ribosomal Intergenic Spacer Analysis*) (Fisher & Triplett 1999). D'une manière générale, ces techniques sont rapides et permettent une analyse comparative simultanée de plusieurs échantillons. Elles ont été mises au point pour observer des différences entre les communautés microbiennes, mais elles ne permettent pas une identification taxonomique directe des communautés.

La PCR quantitative (qPCR) fournit quant à elle une méthode sensible et permettant de quantifier des communautés microbiennes d'intérêt dans des environnements complexes (Zhang & Fang 2006). Elle peut, par exemple, être utilisée pour quantifier, de manière absolue ou relative, les différences observées avec les techniques d'empreintes génétiques.

1.2 Métagénomique et séquençage nouvelle génération

L'analyse des séquences codant pour des biomarqueurs phylogénétiques et fonctionnels est couramment utilisée en écologie microbienne pour l'exploration des environnements complexes. Cependant, même en utilisant des biomarqueurs pour lesquels une multitude d'information est disponible, comme le gène ADNr 16S, ceux-ci ne permettent pas une résolution suffisante dans tous les cas pour assurer une discrimination au niveau de l'espèce ou de la souche (Konstantinidis *et al.* 2006). Les techniques moléculaires assurant l'obtention de l'ensemble des séquences des génomes présents au sein d'un environnement offrent donc une vision plus exhaustive de la diversité génétique (Handelsman *et al.* 1998).

1.2.1 Métagénomique

La métagénomique, également connue sous le terme de génomique environnementale ou génomique des communautés, se définit comme l'étude globale de l'ensemble des génomes des communautés microbiennes multi-espèces extraits directement à partir d'un échantillon environnemental et ne nécessitant pas au préalable une connaissance ou une mise

en culture des communautés microbiennes (Handelsman *et al.* 1998 ; Riesenfeld *et al.* 2004). D'une manière générale, les techniques utilisant la métagénomique sont basées sur le principe suivant : l'ensemble des données génomiques des communautés microbiennes de l'environnement peut être criblé et/ou séquencé de la même manière que la totalité du génome extrait par exemple d'une culture bactérienne pure. Des études métagénomiques ont été conduites au niveau de différents environnements tels que les sols, les lacs, les océans ou encore les drainages miniers acides pour permettre d'avoir accès à la diversité phylogénétique et fonctionnelle d'organismes non cultivés (Tyson *et al.* 2004 ; Handelsman 2004 ; Delmont *et al.* 2011). Ainsi, la métagénomique est primordiale pour la compréhension globale, au sein d'un environnement, des rôles des microorganismes non cultivés et de leurs interactions. Les banques environnementales construites à partir de métagénomes se sont avérées être très utiles pour la découverte de nouveaux gènes d'intérêt, avec des applications potentielles au niveau des biotechnologies, de la médecine et de l'industrie (Steele *et al.* 2009). Ainsi, suite au criblage fonctionnel des banques métagénomiques, des séquences impliquées dans des phénotypes d'intérêt ont été caractérisées. C'est notamment le cas de nouveaux antibiotiques (*e.g.* turbomycine (Gillespie *et al.* 2002), terragine (Wang *et al.* 2000)) (Garmendia *et al.* 2012) ou d'enzymes microbiennes d'intérêt biotechnologique (*e.g.* cellulases, lipases, amylases, nucléases) issues d'environnements variés (Steele *et al.* 2009). Toutefois cette stratégie nécessite l'expression des gènes dans un système hétérologue comme des bactéries ou des levures. Il est donc important de disposer de vecteurs appelés « vecteurs navettes » capables de se propager à la fois dans différents hôtes bactériens et levures (Leis *et al.* 2013). En effet, au niveau d'une banque métagénomique, la fréquence associée à l'identification de gènes actifs exprimant un phénotype d'intérêt est relativement basse lorsqu'un seul type de système hétérologue est utilisé. A titre d'exemple, une étude au niveau d'une banque métagénomique, créée à partir d'un échantillon de sol et utilisant un système hétérologue *Escherichia coli* a pu montrer seulement un clone sur 730 000 possédant une activité lipolytique d'intérêt (Henne *et al.* 2000).

Le criblage du métagénome peut aussi être réalisé *via* l'utilisation de sondes moléculaires ciblant des gènes d'intérêt et chaque fragment d'ADN caractérisé sera alors séquencé (Jacquiod *et al.* 2014). Finalement, l'exploitation des banques métagénomiques peut également se faire par séquençage massif de l'ensemble des fragments générés. Ce type d'approche permet notamment de mettre en lumière d'importantes caractéristiques et organisations au niveau génomique des caractères acquis par des transferts horizontaux de

Tableau 1. Comparaison des différentes plateformes de séquençage de première et deuxième génération.

	Séquenceur (Société)	Méthode d'amplification	Méthode de séquençage	Longueur des lectures	Débit (Mb par run)	Temps de séquençage	Coût (par Mb)	Disponibilité commerciale
1ère génération	**3730xl** (Applied Biosystems by Life Technologies)	PCR, clonage	Séquençage par synthèse (Sanger)	600-1000	0,06	2h	$2 308	Oui
2ème génération	**454 GS Jr. Titanium** (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	400	50	10h	$19,5	Oui
	454 FLX Titanium (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	400	400	10h	$15,5	Oui
	454 FLX+ (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	650	650	20h	$9,5	Oui
	Illumina MiSeq (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	2×300	13 200	55h	$0,1	Oui
	Illumina HiSeq 2500 (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	2×125	500 000	6 jours	$0,03	Oui
	Illumina HiSeq X (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	2×150	1 800 000	3 jours	$0,007	Oui
	SOLiD – 5500xl (Applied Biosystems by Life Technologies)	PCR en émulsion (emPCR)	Séquençage par ligation	110	155 100	8 jours	$0,07	Oui

CRT : *Cyclic Reversible Termination*

gènes (Handelsman 2004) mais également sur le rôle des microorganismes au sein des écosystèmes comme par exemple celui du microbiome humain (Martín *et al.* 2014).

Malgré la nouvelle vision du monde microbien apportée par l'essor de la métagénomique, qui permet d'avoir accès à l'ensemble des microorganismes présents dans un environnement, la création des banques de clones environnementales demeure toutefois techniquement difficile à mettre en œuvre et coûteuse (Rajendhran & Gunasekaran 2008). Actuellement, les nouvelles plateformes de séquençage facilitent l'étude des échantillons en proposant de séquencer directement les acides nucléiques extraits et donc de s'affranchir des étapes de clonage (Shendure & Ji 2008).

1.2.2 La révolution des techniques de séquençage

Grâce au développement de nouvelles technologies de séquençage qui, depuis ces dix dernières années, ont connu une révolution sans précédent, il est désormais possible d'avoir une vision globale et plus intégrative de l'ensemble des évènements se déroulant dans un environnement. Cette révolution concerne à la fois les technologies, l'appareillage, l'informatique ou encore les outils de traitement et de stockage des données (Morey *et al.* 2013).

1.2.2.a De la première à la deuxième génération de séquençage

Au début de la génomique dans les années 1990, le séquençage était réalisé par une première génération basée sur la technique de Sanger (Sanger *et al.* 1977 ; Swerdlow & Gesteland 1990 ; Hunkapiller *et al.* 1991). Même si ce type de séquençage est de moins en moins utilisé du fait de son coût élevé et de son faible débit (**Tableau 1**), il a permis la réalisation de projets d'envergure comme le séquençage du métagénome de la mer des Sargasses (Venter *et al.* 2004) ou encore celui de la surface des océans (intitulé « *the Sorcerer II Global Ocean Sampling Expedition* ») (Rusch *et al.* 2007). Mais depuis ces dix dernières années le séquençage par la méthode de Sanger a laissé place aux techniques de séquençage dites de nouvelle génération ou « *Next Generation Sequencing* » (NGS), qui permettent de s'affranchir des étapes de clonage des fragments d'ADN, de réduire fortement les coûts et le temps d'acquisition des données, et donc d'augmenter considérablement les quantités de données de séquences produites (Shendure & Ji 2008 ; Metzker 2010 ; Zhou *et al.* 2010 ; Shokralla *et al.* 2012 ; Morey *et al.* 2013). Ces nouvelles approches dites à haut-débit sont indispensables pour permettre une exploration fine et présenter une vision non biaisée de la composition phylogénétique et de la diversité fonctionnelle des communautés microbiennes

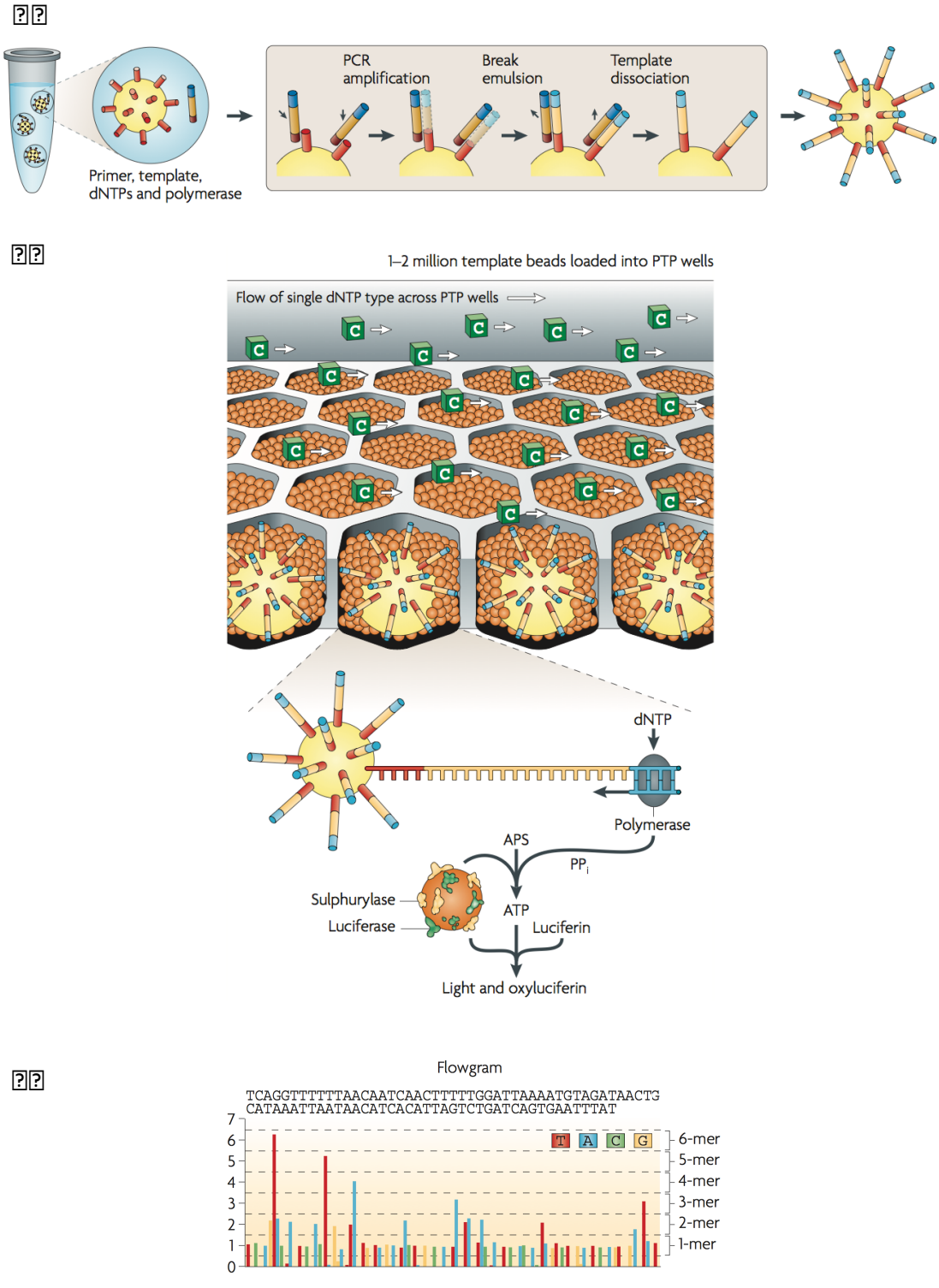


Figure 2. Le pyroséquençage 454.

(A) Les fragments d'ADN à séquencer sont amplifiés à la surface d'une microbille *via* une PCR en émulsion. **(B)** Les microbilles sont ensuite déposées sur une plaque picotitrée (PTP) où sera effectuée la réaction de pyroséquençage. **(C)** L'intensité lumineuse émise est ainsi corrélée au nombre de nucléotides incorporés et l'obtention de la séquence se fait par lecture du *flowgram*. D'après Metzker (2010).

au sein des environnements complexes. Deux principales technologies de séquençage de deuxième génération ont connu un essor considérable en termes de développement technologique et d'application pour les études en écologie microbienne : le pyroséquençage 454 (454 *Life Sciences* / *Roche Applied Science*) et le séquençage Illumina.

iii. Le pyroséquençage 454

La technologie 454 est basée sur le principe du pyroséquençage développé depuis le milieu des années 1980 (Nyrén & Lundin 1985 ; Hyman 1988) puis amélioré au milieu des années 1990 avec la possibilité de multiplexage (Ronaghi *et al.* 1996). Le pyroséquençage est un séquençage par émission de lumière. Les quatre désoxyribonucléotides sont ajoutés un par un de manière itérative et un capteur à transfert de charge (*i.e.* dispositif CDD ou *Charge Coupled Device*) permet de détecter la lumière produite transformant ainsi le signal lumineux en impulsion électrique. Au niveau moléculaire, l'addition au cours de la polymérisation d'un désoxyribonucléotide par l'ADN polymérase aboutit au relargage d'un pyrophosphate inorganique (PPi). Ce même PPi couplé à l'adénosine phosphosulfate est pris en charge par l'ATP sulfurylase pour produire de l'ATP. Enfin, l'ATP néoformé couplé à la D-luciférine aboutit à la production d'oxyluciférine et de lumière par la luciférase. Une apyrase, quant à elle, est chargée de dégrader les désoxyribonucléotides non incorporés et l'ATP résiduel entre deux cycles d'incorporation de bases au niveau du brin néosynthétisé (Ahmadian *et al.* 2006) (**Figure 2**).

La technologie de séquençage de 454 *Life Sciences* / *Roche Applied Science* développée en 2005 (Margulies *et al.* 2005) est issue de la combinaison du pyroséquençage avec l'utilisation de plaques picotitrées (*PicoTiterPlate*, PTP), de la PCR en émulsion (emPCR) ainsi que des technologies informatiques pour l'acquisition et le traitement des images (**Figure 2**). Actuellement, des plateformes comme le GS FLX Titanium permettent de réduire fortement les coûts et le temps d'acquisition des données (**Tableau 1**) (Glenn 2011). Différentes applications de séquençage sont possibles grâce à la technologie 454, avec notamment la possibilité de séquencer à la fois à haut-débit (un million de lectures), de façon globale et sans *a priori*, des métagénomes de manière dite *de novo* (Petrosino *et al.* 2009). Le principal avantage d'utilisation du pyroséquençage 454 demeure la longueur des lectures produites (*i.e.* un million de lectures de 400 à 700 bases suivant la plateforme) avec un temps de séquençage relativement court (*i.e.* entre 10 et 20h) (Glenn 2011). Cette perspective est très intéressante pour des études métagénomiques où l'identification de fragments de grande taille

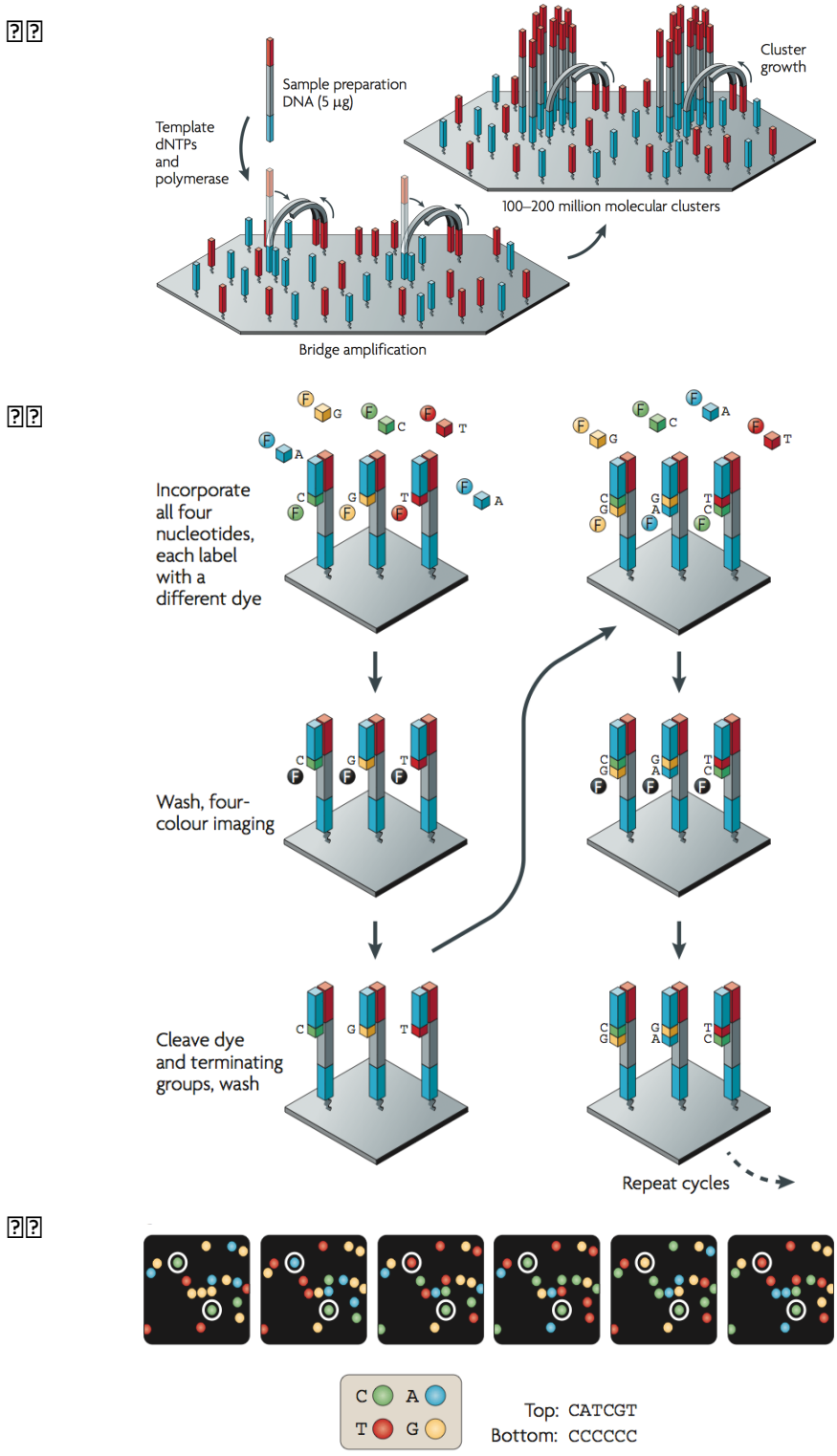


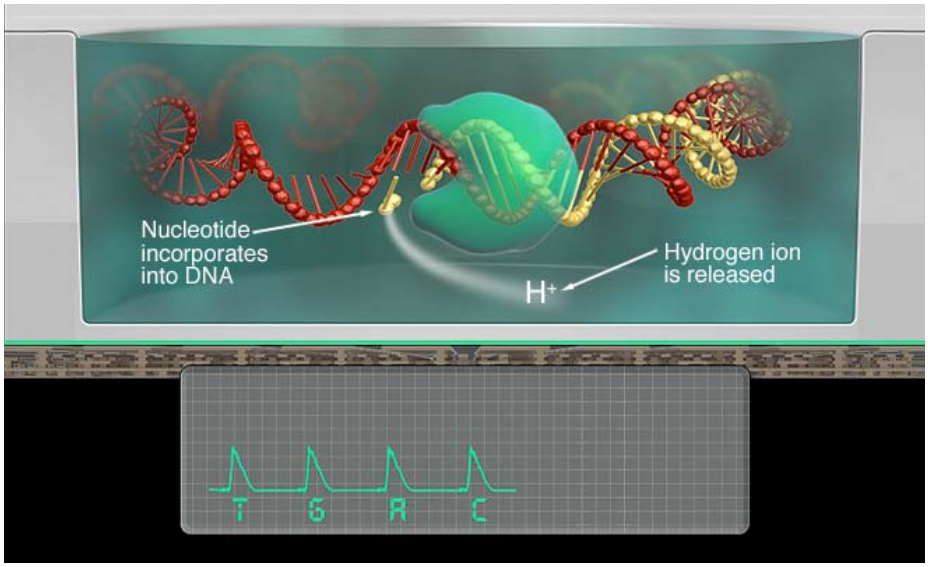
Figure 3. Le séquençage Illumina.

(A) Les fragments d'ADN à séquencer sont amplifiés à la surface d'une *flow cell via* une PCR en ponts. **(B)** Après amplification, le séquençage par terminaison réversible peut avoir lieu. **(C)** A la fin de chaque cycle, la fluorescence émise est interprétée afin d'obtenir la séquence. D'après Metzker (2010).

apporte une information phylogénétique et fonctionnelle plus pertinente (Wommack *et al.* 2008). De plus, cette approche facilite l'assemblage des lectures et permet donc de reconstruire des gènes ou des opérons entiers. Cependant, certaines limites technologiques sont à prendre en considération comme par exemple le coût relativement élevé de cette méthode (environ 10€ la mégabase) et des erreurs de séquençage en relation notamment avec la présence d'homopolymères. En effet, le pyroséquençage n'ayant pas de terminaison pendant la synthèse empêchant l'incorporation multiple de nucléotides au cours d'un cycle de séquençage, ce dernier se base sur l'intensité de la lumière émise pour déterminer le nombre de bases incorporées. Ainsi, le problème fréquemment rencontré est la perte de la relation de linéarité entre l'intensité de lumière émise et le nombre de nucléotides incorporés, aboutissant à des insertions de bases ou des délétions (Margulies *et al.* 2005 ; Rothberg & Leamon 2008 ; Gilles *et al.* 2011). Ces erreurs de séquençage peuvent alors conduire à une représentation biaisée de la diversité (Kunin *et al.* 2010).

iv. Le séquençage Illumina

Cette technologie de séquençage est basée sur le principe de terminaison réversible (*Cyclic Reversible Termination* ou CRT) par utilisation de désoxyribonucléotides triphosphates (dNTPs) modifiés et fluorescents (Metzker 2005 ; Guo *et al.* 2008), et implique une méthode de synthèse en trois étapes : l'incorporation, la mesure de la fluorescence et le clivage (Metzker 2005 ; 2010). Premièrement, une ADN polymérase va initier la synthèse du brin complémentaire au niveau de l'amorce de séquençage en incorporant un dNTP modifié portant un fluorophore et un groupement protecteur au niveau de l'extrémité 3'-OH du ribose. Deuxièmement, suite à l'incorporation, les dNTPs modifiés résiduels seront éliminés par lavage et la fluorescence émise est enregistrée en temps réel permettant de déterminer la nature de la base incorporée au niveau de la séquence. Enfin, l'étape de clivage, suivie d'une nouvelle étape de lavage, élimine le groupement protecteur en 3'-OH inhibant la réaction de polymérisation, ainsi que le fluorophore pour permettre une nouvelle étape d'incorporation (**Figure 3**). Ce principe est dérivé de la méthode de Sanger, où contrairement à cette dernière qui utilise des didésoxyribonucléotides triphosphates (ddNTPs) bloquant la polymérisation, la méthode CRT offre la possibilité de bloquer la polymérisation de manière réversible. La clé de cette méthode réside en l'utilisation de bases bloquantes, comme les 3′-O-azidométhyl-dNTPs, couplées à des fluorophores et portant un groupement azidométhyle sur l'extrémité 3'-OH du ribose, pouvant être clivés chimiquement pour restaurer une extrémité 3'OH libre et ainsi rétablir la polymérisation.



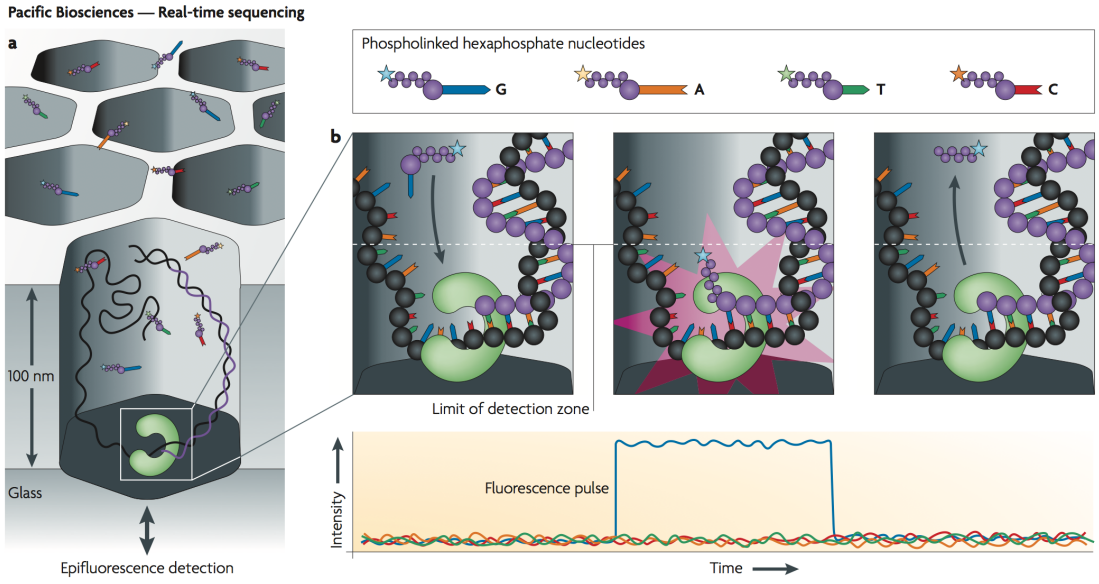


Figure 4. Représentation schématique du mode de fonctionnement des nouvelles technologies de séquençage dites de troisième génération.

(A) Système Ion Torrent où lors de l'incorporation d'un dNTP des ions H^+ sont libérés modifiant le pH à l'intérieur du puits. Ce changement de pH est converti *via* une couche semi-conductrice et une plaque de détection en un signal numérique. D'après http://www.iontorrent.com. **(B)** Système PacBio RS au niveau d'un nanopuit « *Zero Mode Waveguide* » contenant une ADN polymérase fixée au fond du puits. La configuration du ZMW permet une détection uniquement de la fluorescence sur le fond du puits. D'après Metzker (2010).

La technologie de séquençage Illumina (Solexa) est issue de la combinaison de la méthode CRT utilisant quatre fluorophores différents (Turcatti *et al.* 2008), de l'amplification par PCR en ponts sur phase solide (Adessi *et al.* 2000), des nanotechnologies (Fedurco *et al.* 2006) et des technologies informatiques pour l'acquisition et le traitement des images. La technologie Illumina offre un coût de séquençage réduit (Glenn 2011) (**Tableau 1**), avec également la possibilité, comme la technologie 454, de séquencer des métagénomes de manière *de novo* mais à un débit encore plus important (plusieurs milliards de lectures). La technologie Illumina représente ainsi une révolution sans précédent pour les études métagénomiques, qui recherchent une très grande profondeur de séquençage pour étudier la diversité phylogénétique et fonctionnelle des communautés bactériennes (Rodrigue *et al.* 2010 ; Qin *et al.* 2010 ; Hess *et al.* 2011 ; Yu & Zhang 2012). Cependant, le séquençage Illumina possède certaines limites comme le temps de séquençage important (plusieurs jours) mais surtout la longueur des lectures plus faible que pour le pyroséquençage 454 (*i.e.* au maximum 300 bases voire 2×300 bases séquencées à partir des deux extrémités). De plus, des erreurs de séquençage sont observées lors d'une mauvaise incorporation ou d'une mauvaise interprétation de la fluorescence émise (*base calling*), aboutissant à des substitutions de bases. Ces substitutions sont fréquemment rencontrées au niveau des sites moléculaires précédés d'une guanine avec substitution préférentielle d'une adénine par une cytosine (Dohm *et al.* 2008 ; Qu *et al.* 2009 ; Meacham *et al.* 2011 ; Nakamura *et al.* 2011).

1.2.2.b Vers une troisième génération de séquençage

Malgré l'énorme révolution des techniques de séquençage ayant abouti à l'émergence des techniques dites de deuxième génération, le problème majeur à l'heure actuelle est la taille relativement courte des fragments séquencés, compliquant fortement les étapes d'assemblage et donc la reconstruction de génomes complets (Morales & Holben 2011). En outre, malgré la quantité massive de données générées celles-ci restent insuffisantes pour accéder aux génomes de l'ensemble des populations microbiennes présentes au sein des environnements complexes. En effet, Quince *et al.* (2008) ont estimé que pour certains environnements complexes, l'effort de séquençage déployé est insuffisant et doit être multiplié par 10 000 afin de couvrir 90% de la diversité microbienne qu'ils hébergent. Même si les capacités de séquençage ont fortement évolué, cette couverture de la diversité n'est pas envisageable autant sur un plan technologique que financier (Quince *et al.* 2008). D'autres auteurs affirment que pour obtenir un jeu de séquences représentatif d'un gramme de sol avec une

Tableau 2. Comparaison des différentes plateformes de séquençage de troisième génération.

	Séquenceur (Société)	Méthode d'amplification	Méthode de séquençage	Longueur des lectures	Débit (Mb par run)	Temps de séquençage	Coût (par Mb)	Disponibilité
3ème génération	**Ion Torrent – '316'chip** (Life technologies)	PCR en émulsion (emPCR)	Séquençage par synthèse (détection H+)	400	1000	5h	$0,7	Oui
	Ion Torrent – '318'chip (Life technologies)	PCR en émulsion (emPCR)	Séquençage par synthèse (détection H+)	400	1900	7h	$0,5	Oui
	PacBio RS II (Pacific Biosciences)	Aucune	Séquençage de molécules individuelles en temps réel (SMRT)	3000	90	2h	$1,1	Oui
	GridION 8000 (Oxford Nanopore)	Aucune	Séquençage de molécules individuelles (exonucléase et ADN polymérase Phi 29)	10 000	100 000	ND	$0,01	2014 ?
	MinION (Oxford Nanopore)	Aucune	Séquençage de molécules individuelles (exonucléase et ADN polymérase Phi 29)	9000	900	<6h	$1	2014 ?
	Optipore (Noblegen Biosciences)	Aucune	Séquençage de molécules individuelles	ND	ND	ND	ND	2014 ?

ND : Non déterminé
SMRT : *Single Molecule Real Time Technology*

couverture de 1×, il faudrait plus de 6000 *runs* de HiSeq 2000 représentant un coût de séquençage de 267 millions de dollars (Desai *et al.* 2012).

Pour s'affranchir de ces limites, une troisième génération de méthodes de séquençage, exploitant les avancées des nanotechnologies, est apparue (Wash & Image 2008 ; Munroe & Harris 2010 ; Schadt *et al.* 2010b ; Glenn 2011 ; Pareek *et al.* 2011 ; Quail *et al.* 2012 ; Morey *et al.* 2013). Ces nouvelles méthodes sont basées sur l'immobilisation individuelle sur un support solide d'une enzyme (*i.e.* une ADN polymérase ou une exonucléase) permettant de séquencer une seule molécule d'ADN à la fois. Deux nouvelles technologies de séquençage dites de troisième génération sont actuellement disponibles et commercialisées : le système PacBio RS (*Pacific Biosciences*) basé sur le principe de séquençage SMRT (*Single Molecule Real Time Technology)* (Eid *et al.* 2009 ; Korlach *et al.* 2010 ; McCarthy 2010) et le système Ion Torrent (*Life technologies*) basé sur le principe « *Ion semiconductor sequencing* » (Rothberg *et al.* 2011) (**Figure 4**). Ces deux méthodes utilisent un réseau de nanopuits contenant une ADN polymérase permettant la réaction de séquençage. Le système PacBio RS utilise des puits nanophotoniques appelés « *zero-mode waveguide* » (Levene *et al.* 2003) d'un volume de l'ordre du zeptolitre (10^{-21} litre) permettant de canaliser et ainsi d'éviter la propagation de la lumière visible de grande longueur d'onde hors des puits, et donc d'assurer une détection efficace des signaux fluorescents émis lors de l'incorporation des nucléotides par l'ADN polymérase. Le système Ion Torrent n'utilise pas de dNTPs fluorescents, mais un semi-conducteur détectant une différence de potentiel créée par la libération d'ions H^+ suite à l'incorporation d'un dNTP par l'ADN polymérase. Cette technologie nécessite, à la différence du PacBio RS, l'ajout séquentiel des dNTPs. Cette troisième génération de séquençage améliore encore de manière significative les capacités de séquençage (*e.g.* 1,9 Gb en 7h pour Ion Torrent et 90 Mb en 2h pour PacBio RS) avec l'obtention de lectures de grande taille notamment pour le PacBio RS (>3000 bases) (**Tableau 2**) (Glenn 2011). Ces nouvelles techniques ont été récemment appliquées pour le séquençage de génomes complets notamment celui du sérotype O104:H4 d'*Escherichia coli* responsable d'une épidémie par consommation de graines germées en mai 2011 en Allemagne. Le génome complet de la souche a pu être séquencé *via* l'utilisation du système Ion Torrent, puis assemblé en seulement deux jours (Mellmann *et al.* 2011). De même l'utilisation du PacBio RS a permis le séquençage complet des génomes de cinq souches de *Vibrio cholerae*, dont celle responsable de l'épidémie de choléra en Haïti en octobre 2010, et ceci en seulement 3 heures pour permettre des études ultérieures de génomique comparative (Chin *et al.* 2011). Il existe

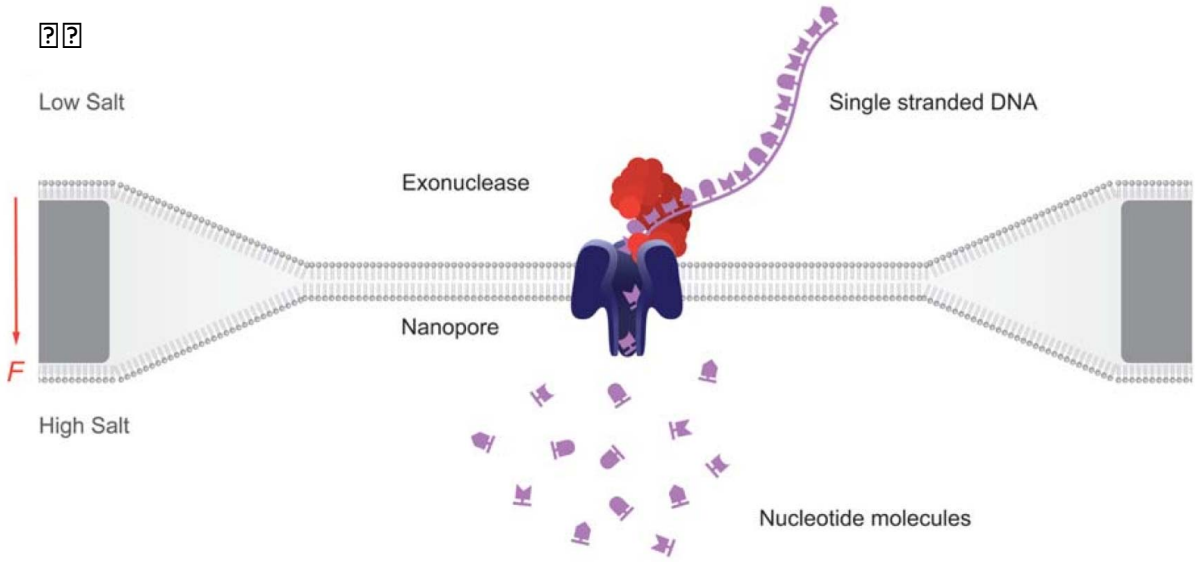


Figure 5. Applications du séquençage de troisième génération de type Nanopore.

(A) Une exonucléase (rouge) fixée sur un nanopore d'α-hemolysine (bleu) dégrade le brin d'ADN en faisant tomber les bases (violet) une par une à travers le nanopore. L'information de séquence est déterminée par une modification de la différence de potentiel à travers le nanopore. D'après Schadt *et al.* (2010). **(B)** Système de séquençage de troisième génération Nanopore MinION prochainement commercialisé sous la forme d'une clé USB. D'après Eisenstein (2012).

également d'autres applications de ces systèmes, en métagénomique par exemple, avec le séquençage Ion Torrent de la communauté microbienne d'un désert salin (Pandit *et al.* 2014), ou des applications du PacBio RS sur des régions génomiques complexes (Huddleston *et al.* 2014) grâce à la longueur des lectures produites.

Toutefois, le domaine du séquençage haut débit est en perpétuelle évolution et d'autres systèmes sont développés pour assurer le séquençage des molécules de grande taille pouvant atteindre une centaine de kilobases. Il est possible de citer la technologie Nanopore (*Oxford Nanopore Technologies*) proposant deux types d'application de séquençage : un système en cours de développement dans lequel une exonucléase est fixée sur un nanopore d'α-hémolysine (Clarke *et al.* 2009 ; Timp *et al.* 2010) et un système en voie de commercialisation utilisant un nanopore à base d'une porine A de *Mycobacterium smegmatis* (MspA) et une ADN polymérase *phi29* (Cherf *et al.* 2012 ; Manrao *et al.* 2012 ; Schneider & Dekker 2012). Ce système s'utilisera au sein de deux plateformes : l'une appelée GridION et l'autre MinION correspondant à un système miniature de séquençage, se présentant sous la forme d'une clé USB, pouvant être relié à un ordinateur portable (Eisenstein 2012 ; Laszlo *et al.* 2014) (**Figure 5**). Une autre application des nanopores est exploitée par la société Noblegen Biosciences qui ambitionne de commercialiser prochainement le système « optipore » (pour *optical detection* et *nanopore*), un séquenceur de paillasse de troisième génération combinant les nanotechnologies et un système de lecture optique (McNally *et al.* 2010 ; Singer *et al.* 2012).

Ces technologies émergentes sont très prometteuses pour des applications en écologie microbienne et plus précisément pour l'étude des environnements complexes grâce notamment à la longueur des lectures générées. Néanmoins, leurs applications restent encore limitées pour les échantillons métagénomiques du fait d'un taux d'erreur de séquençage relativement important (jusqu'à 16%) pour le système PacBio RS (Glenn 2011).

L'utilisation des nouvelles techniques de séquençage à l'échelle du laboratoire démeure également problématique en raison de la masse de données produites restant très délicate à analyser. Cette perspective implique obligatoirement la collaboration de plusieurs équipes de recherche, de disposer de moyens de calcul, de traitement et de stockage conséquents.

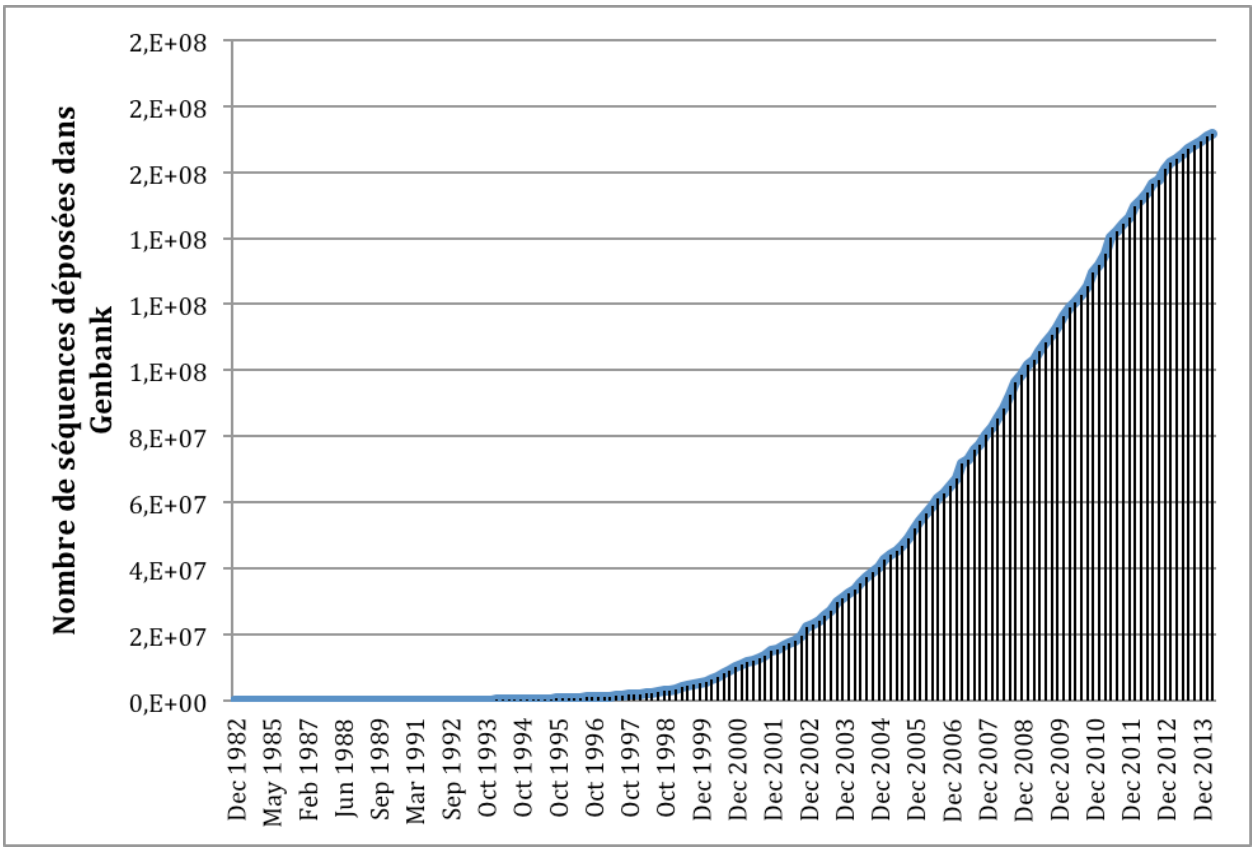


Figure 6. Croissance exponentielle des données de séquences disponibles.

Le graphique représente le nombre de séquences répertoriées dans GenBank depuis la création de cette base de données. Le temps moyen de doublement du nombre de séquences est aujourd'hui estimé à 18 mois.

1.3 Défis et limites bioinformatiques

La recherche en génomique environnementale est en train de vivre une véritable révolution de l'information avec la possibilité d'accéder rapidement et de façon exhaustive aux séquences génomiques. Des jeux de données de plus en plus importants, pour un nombre croissant d'organismes mais également d'écosystèmes, sont mis à disposition de la communauté scientifique. Toutefois l'avalanche de ces données est telle que l'on se heurte aujourd'hui aux difficultés concernant leur stockage, leur partage ou leur analyse (Pelletier & Perrière 2013).

1.3.1 Stockage, accès et partage des données de séquençage haut-débit

Depuis près de 30 ans, les trois grandes banques généralistes de données collectant les séquences génomiques sont : GenBank au *National Center for Biotechnology Information* (Benson *et al.* 2014), l'ENA (*European Nucleotide Archive*) à l'*European Bioinformatics Institute* (Brooksbank *et al.* 2014) et la DDBJ (*DNA Data Bank of Japan*) au *National Institute of Genetics* (Kosuge *et al.* 2014). Bien qu'à leurs débuts le contenu et la taille de ces trois banques étaient relativement différents, une collaboration internationale s'est rapidement établie et, depuis 25 ans, leurs contenus sont virtuellement identiques. Ces bases de données permettent ainsi d'accéder librement à la quasi-totalité des séquences biologiques obtenues par la communauté scientifique.

Avec l'avènement des méthodes de séquençage durant les dernières décennies, le volume des données soumises à ces trois bases de données généralistes a cru de façon exponentielle, avec un temps de doublement moyen de l'ordre de 18 mois (**Figure 6**). De 2000 à 2010 le temps de doublement était plus faible en raison du séquençage de nombreux génomes ou transcriptomes, dont le premier génome humain (Venter *et al.* 2001). Cependant, ces dernières années, l'afflux de données semble se stabiliser alors que les méthodes de séquençage à très haut débit se sont généralisées et qu'il n'y a jamais eu autant de séquences produites dans les laboratoires. Une première explication à ce phénomène inattendu tient au fait que les centres en charge de la maintenance des banques sont de moins en moins à même de supporter les charges financières que représente l'achat continuel de capacités de stockage supplémentaires ainsi que la maintenance des infrastructures associées (Pelletier & Perrière 2013).

Un autre problème vient du volume de données produites qui est désormais tel qu'il n'est plus possible de les transmettre en un temps raisonnable *via* le réseau. En effet, un *run*

complet d'Illumina HiSeq 2500 produit près de 1 To de données brutes et la solution de plus en plus utilisée pour leur transfert est l'expédition d'un disque dur sur lequel sont sauvegardées les séquences.

Enfin, la question de l'accès aux lectures courtes, non annotées, est également un problème d'importance. En raison du nombre très important de séquences les bases de données généralistes ne proposent plus un accès direct aux séquences individuelles, mais plutôt à des archives compressées contenant l'ensemble des lectures d'un *run*. C'est notamment le cas de la base de données SRA (« *Short Read Archive* ») (Shumway *et al.* 2010) du NCBI qui contient plus de 1000 Terabases (24/07/2014). Dans ce contexte, de plus en plus de séquences ne sont plus envoyées à l'un des trois grands centres de saisie mais sont mises à disposition de la communauté par l'intermédiaire de banques de données locales mises en place dans le cadre de projets limités. Il existe ainsi de très nombreuses bases de données spécialisées, que celles-ci soient dédiées à un organisme ou une problématique biologique particulière (NCBI Resource Coordinators 2014). Les trois collections généralistes ne peuvent donc plus être considérées comme exhaustives et ceci a d'ores et déjà des répercussions sur la reproductibilité des résultats. C'est dans le but de pallier ce problème, que l'EBI a lancé, début 2007, l'initiative ELIXIR. Cette initiative vise à fédérer les grands centres de bioinformatique (nationaux ou régionaux) dans un réseau Européen, chaque nœud ayant une spécificité thématique. Une telle réorganisation permettrait effectivement de répartir la charge, aussi bien en quantité de données à gérer qu'en termes d'infrastructures, pour ce qui est du stockage ou de l'archivage des séquences biologiques (Pelletier & Perrière 2013).

1.3.2 Méthodes d'analyses des séquences métagénomiques

Les flux d'acquisition de données deviennent extrêmement rapides et volumineux ce qui pose le problème de leur gestion, stockage, mais aussi de leur exploitation.

Le développement des plateformes de séquençage de deuxième génération a conduit à la production de données de séquences à des coûts très bas avec des débits considérables (Glenn 2011). Cependant, ces évolutions se sont faites au détriment de la longueur et de la qualité des séquences posant de nouveaux problèmes d'analyse. Ce déluge de séquences nécessite donc le développement de nouveaux outils bioinformatiques pour assurer un traitement optimal de l'information (Wooley *et al.* 2010 ; Logares *et al.* 2012).

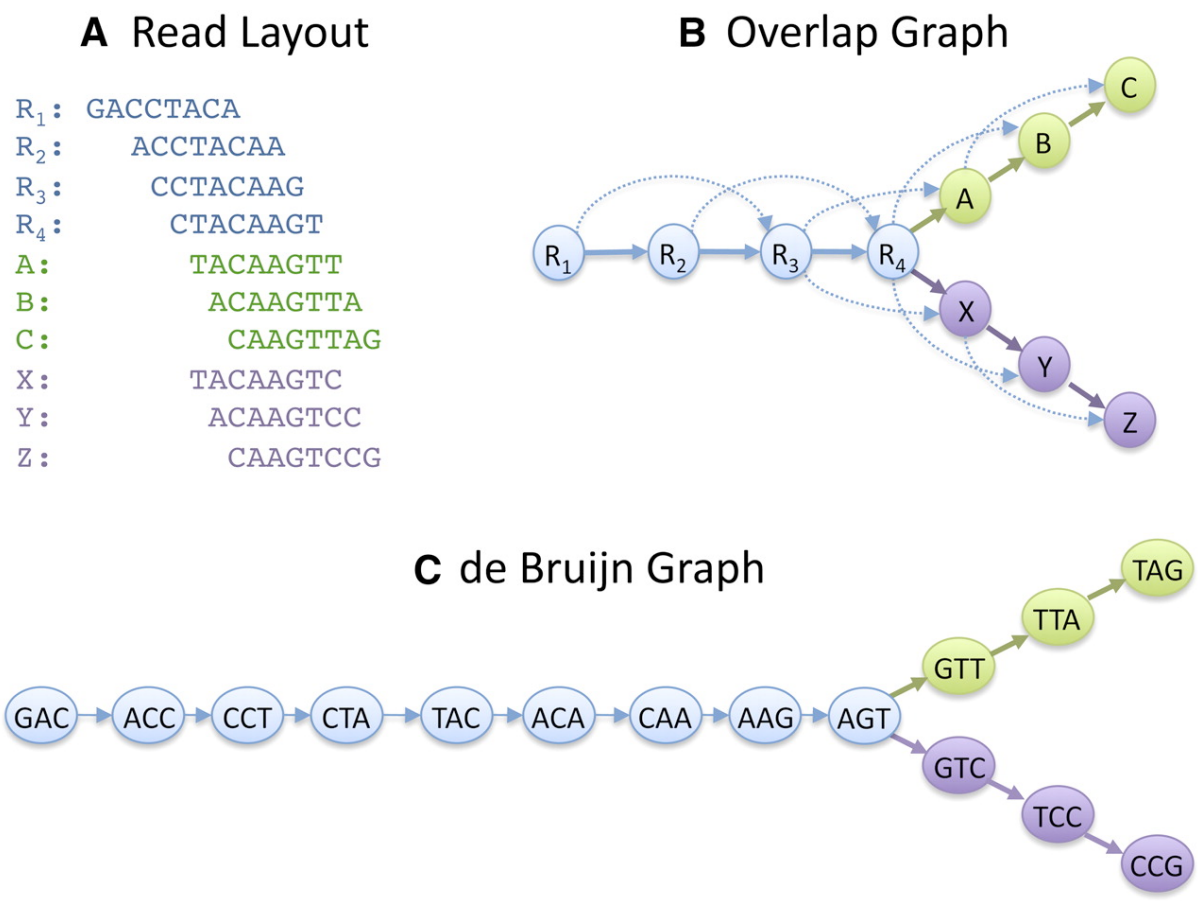


Figure 7. Assemblage des données de séquençage haut-débit *via* l'utilisation de graphes. **(A)** A partir de 10 lectures de 8 nucléotides, il est possible de construire **(B)** un graphe de chevauchement (*Overlap graph*) où chaque lecture est un nœud et une arête relie deux nœuds lorsque ceux-ci présentent un chevauchement d'au moins 5 nucléotides. **(C)** Dans un graphe de De Bruijn un nœud est créé pour chaque *k*-mer issu de la totalité des lectures. Une arête relie deux nœuds si les *k*-mers se chevauchent sur *k*-1 nucléotides. D'après Schatz *et al.* (2010).

1.3.2.a Qualité des données

La qualité des données issues des nouvelles techniques de séquençage est liée non seulement à la technologie utilisée mais également, dans certains cas, aux étapes préliminaires ayant permis l'obtention du matériel génétique à séquencer (*e.g.* échantillonnage, extraction des acides nucléiques) (Peyret 2013). Ainsi, la première étape de l'analyse bioinformatique des données doit s'attacher à détecter les régions de mauvaise qualité, identifier les erreurs de séquençage ainsi que les séquences issues d'artéfacts de manipulation (Quince *et al.* 2011). En effet, des séquences de mauvaise qualité et/ou artéfactuelles peuvent compromettre les analyses ultérieures comme l'assemblage ou l'annotation mais également surestimer une diversité non représentative des environnements explorés (Kunin *et al.* 2010 ; Bachy *et al.* 2013). Au final, seules les séquences de qualité qui auront été retenues permettront de refléter le plus fidèlement possible l'information génétique initiale issue des échantillons rendant ainsi possibles des traitements statistiques afin de tester les hypothèses initialement posées (Peyret 2013).

1.3.2.b Assemblage de métagénomes

L'assemblage des données de séquençage haut-débit a nécessité une refonte complète des algorithmes. Initialement, l'assemblage des données de séquençage de première génération (Sanger) était basé sur l'alignement de toutes les paires de séquences en faisant l'hypothèse que toutes les séquences appartiennent au même organisme. La métagénomique, en apportant une information de séquence correspondant jusqu'à plusieurs milliers d'organismes, a donc profondément bouleversé les méthodes d'assemblage.

Aujourd'hui, l'assemblage *de novo* d'un métagénome ne repose donc pas sur un algorithme « glouton » (*greedy algorithm*) mais principalement sur la théorie des graphes (Miller *et al.* 2010 ; Nagarajan & Pop 2013). Ainsi, chaque lecture est découpée en mots de *k* nucléotides (*k*-mers), puis l'assembleur construit un graphe orienté représentant tous les chevauchements de longueur *k*-1 entre tous les *k*-mers. Chaque nœud est représenté par un *k*-mer, et deux nœuds sont reliés par une arête si ils sont chevauchants sur *k*-1 nucléotides, on parle alors de graphe de De Bruijn (**Figure 7**). Chaque *contig* résultant de la fusion de plusieurs lectures chevauchantes correspond ainsi à un chemin dans ce graphe (Schatz *et al.* 2010 ; Nagarajan & Pop 2013).

Cependant, la nature même des données métagénomiques rend leur assemblage délicat. En effet, la représentation inégale des organismes au sein d'un échantillon

Tableau 3. Liste des algorithmes de classification taxonomique des données métagénomiques.

Logiciel	Référence	Composition	Similarité
AbundanceBin	(Wu & Ye 2011)	✔	
AMPHORA	(Wu & Eisen 2008)		✔
BLAST	(Altschul *et al*. 1990)		✔
C16S	(Ghosh *et al*. 2012)		✔
CAMERA	(Seshadri *et al*. 2007)		✔
CARMA	(Gerlach & Stoye 2011)		✔
ClaMS	(Pati *et al*. 2011)	✔	
CloudLCA	(Zhao *et al*. 2012)		✔
CompostBin	(Chatterji *et al*. 2008)	✔	
DiScRIBinATE	(Ghosh *et al*. 2010)		✔
EMMSA	(Kotamarti *et al*. 2010)	✔	
EPA	(Berger & Stamatakis 2011)		✔
ε-NB	(Parks *et al*. 2011)	✔	✔
Eu-Detect	(Mohammed *et al*. 2011a)	✔	
FACS	(Stranneheim *et al*. 2010)		✔
Genometa	(Davenport *et al*. 2012)		✔
GRAMMy	(Xia *et al*. 2011)	✔	
GSMer	(Tu *et al*. 2014)	✔	✔
GSTaxClassifier	(Yu *et al*. 2010)	✔	
H^2SOM	(Martin *et al*. 2008)	✔	
INDUS	(Mohammed *et al*. 2011b)	✔	
Kraken	(Wood & Salzberg 2014)	✔	✔
LikelyBin	(Kislyuk *et al*. 2009)	✔	
MARTA	(Horton *et al*. 2010)		✔
MEGAN	(Huson *et al*. 2007)		✔
MetaBin	(Sharma *et al*. 2012)		✔
MetaCluster-TA	(Wang *et al*. 2014)	✔	✔
MetaCV	(Liu *et al*. 2012)	✔	
MetaID	(Srinivasan & Guda 2013)	✔	
MetaPhlan	(Segata *et al*. 2012)		✔
Metaphyl	(Tanaseichuk *et al*. 2013)		✔
MetaPhyler	(Liu *et al*. 2010)		✔
MetaSAMS	(Zakrzewski *et al*. 2012)	✔	✔
MG-DOTUR	(Schloss & Handelsman 2008)		✔
MLTreeMap	(Stark *et al*. 2010)		✔
mOTU-LGs	(Sunagawa *et al*. 2013)		✔
MG-RAST	(Meyer *et al*. 2008)		✔
MGTAXA	(McLean *et al*. 2013)	✔	
MTR	(Gori *et al*. 2011)		✔
MyTaxa	(Luo *et al*. 2014)		✔
NBC	(Rosen *et al*. 2011)	✔	
PANAM	(Taib *et al*. 2013)		✔
Parallel-META 2.0	(Su *et al*. 2014)		✔
PhyloPythia	(McHardy *et al*. 2007)	✔	
PhyloPythiaS	(Patil *et al*. 2012)	✔	
PhyloSift	(Darling *et al*. 2014)		✔
PhylOTU	(Sharpton *et al*. 2011)		✔
Phymm	(Brady & Salzberg 2009)	✔	
PhymmBL	(Brady & Salzberg 2009)	✔	✔
pplacer	(Matsen *et al*. 2010)		✔
ProViDE	(Ghosh *et al*. 2011)		✔
QIIME	(Caporaso *et al*. 2010)	✔	✔
RAIphy	(Nalbantoglu *et al*. 2011)	✔	
RDP Classifier	(Wang *et al*. 2007)	✔	
RITA	(Macdonald *et al*. 2012)		✔
S-GSOM	(Chan *et al*. 2008)	✔	
SAP	(Munch *et al*. 2008)		✔
Scimm	(Kelley & Salzberg 2010)	✔	
SEK	(Chatterjee *et al*. 2014)	✔	
SIMCOMP	(Prabhakara & Acharya 2010)	✔	✔
SOM	(Abe *et al*. 2005)	✔	
SOrt-ITEMS	(Monzoorul Haque *et al*. 2009)		✔
SPANNER	(Porter & Beiko 2013)		✔
SPHINX	(Mohammed *et al*. 2011c)	✔	✔
SSuMMo	(Leach *et al*. 2012)		✔
STAP	(Wu *et al*. 2008)		✔
TACOA	(Diaz *et al*. 2009)	✔	
TANGO	(Alonso-Alemany *et al*. 2013)		✔
Taxator-tk	(Dröge *et al*. 2014)		✔
TaxSOM	(Weber *et al*. 2011)	✔	
Taxy	(Meinicke *et al*. 2011)	✔	
TETRA	(Teeling *et al*. 2004)	✔	
Treephyler	(Schreiber *et al*. 2010)		✔
TUIT	(Tuzhikov *et al*. 2014)		✔
TWARIT	(Rachamalla *et al*. 2012)	✔	✔
WATERS	(Hartman *et al*. 2010)		✔
WGSQuikr	(Koslicki *et al*. 2014)	✔	

métagénomique ou le polymorphisme de séquences au sein d'individus appartenant à une même espèce entrainent généralement la construction de très nombreux *contigs*, de faible taille et indépendants qui couvrent des régions génomiques des espèces les plus abondantes (Charuvaka & Rangwala 2011). De plus, la présence des séquences identiques entre des espèces taxonomiquement proches peut conduire à l'obtention de *contigs* chimériques. Ainsi, l'arrivée des technologies de séquençage de troisième génération, en produisant des fragments de plus grandes tailles, permettra d'améliorer sensiblement l'assemblage des données métagénomiques (Wooley *et al.* 2010).

1.3.2.c Annotation des données métagénomiques

Une étape cruciale en métagénomique couplée à du séquençage haut-débit est l'annotation taxonomique et fonctionnelle des séquences obtenues afin d'évaluer la structure de la communauté microbienne et les potentialités métaboliques au sein de l'environnement étudié. Or, la nature fragmentaire des données métagénomiques couplée à l'extraordinaire diversité microbienne rend difficile cette étape d'annotation (Simon & Daniel 2011). En effet, les lectures sont souvent trop courtes pour contenir un biomarqueur phylogénétique dans son intégralité ou un cadre de lecture ouvert (*Open Reading Frame* ou ORF) entier. L'annotation taxonomique ou fonctionnelle de ces données représente donc un défi à part entière (Wooley & Ye 2009).

i. Affiliation taxonomique

On peut distinguer deux méthodes principales pour l'affiliation taxonomique des séquences métagénomiques : les méthodes basées sur la similarité de séquences et celles basées sur la composition des séquences (Bazinet & Cummings 2012 ; Mande *et al.* 2012 ; Dröge & McHardy 2012) (**Tableau 3**).

La procédure basée sur la similarité repose sur la recherche d'homologues, dans les bases de données généralistes, par l'intermédiaire d'un outil tel que BLAST (Altschul *et al.* 1990). Alors que certaines méthodes s'arrêtent après cette étape en proposant une affiliation sur la base du (ou des) meilleur(s) résultat(s) BLAST, d'autres algorithmes sélectionnent un ensemble de candidats en fonction de la qualité d'alignement avec la séquence requête pour construire un alignement multiple. Finalement, en appliquant une méthode de reconstruction phylogénétique, cet alignement permet l'élaboration d'un arbre garantissant l'affiliation taxonomique de la séquence étudiée. Néanmoins, cette méthode ne permet pas l'affiliation des

séquences correspondant à de nouvelles espèces encore non identifiées ne possédant donc pas d'homologues proches dans les bases actuelles.

Les méthodes d'affiliation taxonomique basées sur la composition des séquences permettent de s'affranchir de cette limite puisqu'elles reposent sur l'analyse des caractéristiques intrinsèques des séquences. En effet, de nombreux mécanismes comme la réplication, la recombinaison, la réparation de l'ADN, les systèmes de modification par enzymes de restrictions ou la structure de l'ADN (Karlin *et al.* 1997) sont sources de « signatures » génomiques propres à chaque génome. Ainsi, l'étude de la composition des séquences, que ce soit par le contenu en bases Guanine (G) et Cytosine (C) ou par la fréquence de motifs courts (moins de 10 nucléotides), permet la discrimination rapide des séquences pour assurer leur classification. Les algorithmes associés sont basés sur des méthodes statistiques de reconnaissance de ces signatures, comme les modèles de Markov, pour construire des classifieurs automatisés. On distingue alors les approches supervisées qui vont comparer les fréquences des signatures avec des séquences de référence dont l'affiliation est connue pour entraîner leurs classifieurs, et les approches non-supervisées qui utilisent le jeu de séquences en cours d'affiliation pour l'entrainement du classifieur (Bazinet & Cummings 2012 ; Mande *et al.* 2012 ; Dröge & McHardy 2012).

La taille des séquences demeure un paramètre critique pour toutes les méthodes basées sur la composition des séquences. En effet, aucune d'entre elles ne fonctionne réellement efficacement sur des séquences de moins de 1 kpb en raison du nombre limité de mots qu'elles contiennent (et de la variation locale de composition le long du génome) (McHardy & Rigoutsos 2007). Par ailleurs, cette famille de méthodes est très sensible aux erreurs de séquençage ainsi qu'aux transferts horizontaux de gènes. Pour contourner les limites de ces deux grandes familles d'algorithmes, des méthodes hybrides ont récemment été développées en associant les deux méthodes (Bazinet & Cummings 2012).

ii. Annotation fonctionnelle

L'annotation fonctionnelle des données métagénomiques repose généralement sur la détermination préalable des régions codantes (De Filippo *et al.* 2012). Néanmoins, cette prédiction peut être rendue difficile par la taille des séquences et la présence d'erreurs de séquençage. En effet, la plupart des algorithmes d'annotation syntaxique reposent sur l'identification d'ORFs. Or, des erreurs de type insertions/délétions peuvent entraîner un décalage au sein de ces cadres de lecture (*frameshift*), alors que des erreurs de prédictions de

base peuvent l'interrompre par l'introduction d'un codon de terminaison. Des algorithmes de prédictions de gènes dans les données métagénomiques ont donc été développés afin de détecter des régions codantes à partir de lectures courtes et en s'affranchissant de potentielles erreurs de séquençage (Rho *et al.* 2010 ; Kelley *et al.* 2011).

A l'image de l'annotation taxonomique, il existe deux stratégies principales pour prédire la fonction des régions codantes précédemment identifiées : les méthodes basées sur la similarité de séquences et celles basées sur l'identification de domaines ou motifs conservés. Ainsi, la première consiste à comparer la séquence requête à de multiples bases de données, aussi bien généralistes comme Genbank (Benson *et al.* 2014) que spécialisées comme des bases de données d'orthologues (*e.g.* COG (Tatusov *et al.* 2003)), de familles de protéines (*e.g.* Pfam (Finn *et al.* 2014)) ou métaboliques (*e.g.* KEGG (Kanehisa *et al.* 2014)). *A contrario* les méthodes basées sur l'identification de motifs vont chercher à identifier des domaines protéiques conservés, de fonction connue, et répertoriés dans des bases de données spécialisées comme InterPro (Hunter *et al.* 2012) ou PROSITE (Sigrist *et al.* 2013).

Face à la multiplication des outils et stratégies pour l'analyse des séquences métagénomiques, des services web entièrement automatisés ont vu le jour en proposant une prise en charge complète des données. Depuis l'analyse qualité des séquences jusqu'à la reconstruction de voies métaboliques, en passant par la description taxonomique des communautés étudiées, des *pipelines* comme MG-RAST (Meyer *et al.* 2008) ou EBI Metagenomics (Hunter *et al.* 2013) sont de plus en plus utilisés.

1.3.3 Ressources de calcul pour l'analyse des données de séquençage haut-débit

Face à l'augmentation constante des débits de séquençage, les besoins en capacités de calculs se trouvent démultipliés conduisant au déploiement de nouvelles infrastructures informatiques. Ces besoins pour le traitement des données massives de séquençage relèvent aussi bien du calcul intensif (*High Performance Computing* ou HPC) que du traitement à haut-débit (*High Throughput Computing* ou HTC) (Schadt *et al.* 2010a). Le calcul intensif comme l'assemblage des données peut s'effectuer sur des machines multiprocesseurs (*Symmetric MultiProcessing* ou SMP), des *clusters* de calculs (*i.e.* plusieurs machines reliées entre elles *via* un réseau local) ou des architectures de type GPU (*Graphics Processing Unit*) qui tirent profit de la puissance des cartes graphiques. Le traitement à haut-débit d'un très grand nombre de tâches (*e.g.* annotation taxonomique ou fonctionnelle) utilise des

architectures différentes comme les grilles informatiques (*i.e.* plusieurs *clusters* géographiquement distants reliés *via* le réseau Internet). Enfin, récemment une nouvelle architecture de calcul tend à se démocratiser, il s'agit du *cloud computing*. Basé sur la virtualisation, le *cloud computing* repose sur la location temporaire de ressources auprès de compagnies comme Amazon.

Néanmoins, le défi majeur de l'exploitation des données de séquençage à haut-débit est de fournir aux biologistes la capacité de lancer des chaines de traitement sur ces supercalculateurs, grilles ou *clouds* selon leurs besoins (Peyret 2013). En effet, on estime à l'heure actuelle qu'un chercheur occupe 25% de son temps à produire des données et 75% à les analyser. Il est donc nécessaire de se tourner vers l'utilisation ou le développement d'approches moléculaires alternatives permettant d'aborder les problématiques d'écologie microbienne de manière plus ciblée. Par opposition au séquençage direct et systématique des environnements, ces approches de réduction de complexité s'accompagneraient alors de temps de traitement des données beaucoup plus courts.

2. Méthodes de réduction ciblée de la complexité

La métagénomique couplée au séquençage haut-débit a amélioré notre vision de la diversité microbienne. La masse de séquences produite a permis la reconstruction de génomes complets, notamment pour les microorganismes les plus abondants. Elle a également mis en évidence des populations présentes en faible abondance, difficilement détectables sans une profondeur de séquençage importante, mais ayant potentiellement un rôle écologique prépondérant (Hugoni *et al.* 2013). Cette « biosphère rare », longtemps sous-estimée et décriée, commence aujourd'hui à être de plus en plus étudiée (Sogin *et al.* 2006 ; Pedrós-Alió 2007 ; 2012). Néanmoins, étudier la biosphère rare *via* le séquençage massif des écosystèmes impliquerait des milliers de *runs*, plusieurs millions de dollars et des moyens humains et informatiques démesurés (Desai *et al.* 2012). Une alternative intéressante est donc de pouvoir réduire la complexité des échantillons en isolant des individus ou en ciblant, de manière spécifique des séquences nucléiques d'intérêt pour permettre l'application des NGS sur ces mêmes séquences ciblées.

2.1 Amplicons, cellule isolée et capture de gènes

L'avènement des méthodes moléculaires en écologie microbienne a permis le développement de nouvelles stratégies pour analyser de manière ciblée les environnements complexes (Suenaga 2011). La plus ancienne et la plus familière est sans doute la PCR qui utilise un couple d'amorces pour amplifier de manière spécifique une région génomique. Comme pour n'importe quelle molécule d'ADN, les produits PCR peuvent être utilisés comme matrice pour le séquençage utilisant les NGS. Cette approche qualifiée de métagénomique dirigée, métagénétique ou *metabarcoding* a été appliquée à de nombreux biomarqueurs aussi bien phylogénétiques que fonctionnels (Suenaga 2011). Néanmoins, bien que cette méthode ait été intensément utilisée, elle ne permet pas de répondre à la problématique principale en écologie microbienne qui vise à relier structure et fonction des communautés (Morales & Holben 2011 ; Yoccoz 2012). De plus, les nombreux biais inhérents à l'extraction du matériel génétique ou à la PCR peuvent aboutir à une vision erronée de la diversité phylogénétique et fonctionnelle des environnements étudiés (Schloss *et al.* 2011 ; Schloss & Westcott 2011 ; Patin *et al.* 2012).

Les limites de l'amplification PCR peuvent être classées en deux catégories. Elles peuvent être liées à des erreurs de PCR produisant des séquences artéfactuelles, ou résulter d'une efficacité d'amplification non homogène faussant la distribution des amplicons

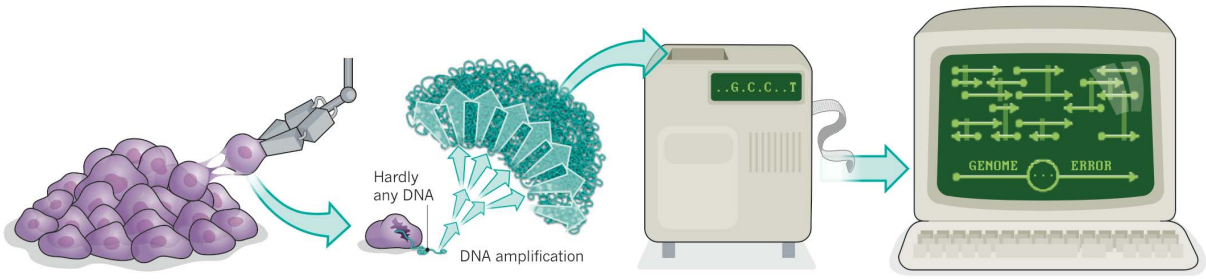


Figure 8. Approche de séquençage par cellule isolée.

La première étape consiste en l'isolement des cellules par micromanipulation, dilution ou tri automatisé. Après leur lyse et l'extraction de leur matériel génétique, une amplification pangénomique de l'ADN est réalisée préalablement au séquençage. D'après Owens (2012).

(Wintzingerode *et al.* 1997 ; Polz & Cavanaugh 1998 ; Acinas *et al.* 2005). Ainsi, les artéfacts de séquences peuvent être dus soit i) à la formation de chimères se produisant au niveau d'un cycle de PCR lors d'une extension incomplète des amorces ; ii) à la formation d'hétéroduplexes (séquences amplicons hétérologues) ; iii) aux erreurs de Taq polymérase où une faible fidélité de cette dernière génère des erreurs d'incorporation durant la synthèse du brin d'ADN aboutissant à des substitutions de base (Wintzingerode *et al.* 1997). Ces séquences artéfactuelles se révèlent problématiques en écologie microbienne, car elles aboutissent de manière erronée à une surestimation de la diversité présente dans un environnement et à une identification de nouveaux variants génétiques. De plus, la présence de bases incorrectement incorporées au cours de l'élongation du brin d'ADN, induite par une faible fidélité de la Taq polymérase, peut être problématique notamment lorsque ces bases sont situées au niveau de sites moléculaires d'intérêt comme par exemple ceux nécessaires à la détermination de sondes ou d'amorces. L'amplification PCR non homogène, quant à elle, peut être due soit i) à l'inhibition de l'amplification par la présence de molécules co-extraites avec les acides nucléiques comme par exemple les acides humiques, ii) à la composition en bases des gènes cibles en lien avec le pourcentage GC des séquences ou encore iii) à la spécificité des amorces utilisées pouvant conduire à une surreprésentation de certains fragments (Suzuki & Giovannoni 1996 ; Wintzingerode *et al.* 1997 ; Polz & Cavanaugh 1998).

Un des défis majeurs actuels en écologie microbienne est donc de pouvoir explorer les environnements complexes en isolant des gènes, des opérons voire des génomes tout en s'affranchissant de la PCR. De nouvelles méthodes moléculaires de réduction de complexité proposent de lever ce verrou technique pour isoler des génomes ou enrichir spécifiquement de grandes régions génomiques d'intérêt. Parmi elles, il est possible de citer les techniques de cellules isolées (*single cell*) qui permettent le séquençage de génomes à partir d'une seule cellule (Raghunathan *et al.* 2005 ; Lasken 2007 ; Ishoey *et al.* 2008 ; Lasken 2012 ; Owens 2012 ; Blainey 2013 ; Nawy 2014) (**Figure 8**). Après l'isolement des cellules (soit par micromanipulation, dilution ou tri automatisé) et leur lyse, une amplification pangénomique de l'ADN (*Whole Genome Amplification*) est réalisée, le plus souvent par MDA (*Multiple Displacement Amplification*) (Dean *et al.* 2001 ; 2002), afin de construire des banques pouvant être séquencées par NGS. Cette stratégie, en cours de développement, a déjà permis de reconstruire partiellement de nombreux génomes notamment d'organismes pour lesquels les autres méthodes moléculaires sont difficilement applicables. Ainsi, Rinke et collaborateurs

(Rinke *et al.* 2013) ont pu résoudre certaines branches de l'arbre du vivant et identifier de nouveaux *phyla* et propriétés métaboliques. Néanmoins, cette approche souffre de plusieurs limites. Parmi celles-ci, on retrouve les problèmes de lyse, les biais engendrés par l'amplification MDA (*i.e.* formation de chimères, amplification non homogène) et les diverses contaminations lors de l'isolement des cellules ou l'amplification (Lasken 2012 ; Blainey 2013).

Afin d'avoir une vue globale des méthodes moléculaires d'enrichissement applicables aux problématiques d'écologie microbienne, nous avons réalisé un état de l'art à partir de données bibliographiques. Cette partie est présentée sous forme d'une revue scientifique, qui sera soumise dans le journal « *Environmental Microbiology Reports* », et offre une présentation détaillée des nouvelles approches d'enrichissement.

Article n°1

Capturing microbial dark matter to illuminate ecosystem functioning

Capturing microbial dark matter to illuminate ecosystem functioning

Abstract

Microbial communities show the greatest organism diversity on earth and are crucial for ecosystem functioning. Culture-independent molecular approaches, targeting mainly small subunit ribosomal RNA genes, have revealed this extraordinary diversity, particularly through the discovery of candidate divisions. Next generation sequencing technologies greatly improved the resolution for microbial diversity description. Thus, the rare biosphere now becomes visible and complete genomes are reconstructed, linking microbial communities structure and realized metabolic functions. Nevertheless, even with the current high throughput sequencing technologies, complete genome assembly is only achievable for dominant micro-organisms in explored ecosystems. A promising approach, based on single cell sequencing strategy, has been developed to overcome genome reconstruction limitations from metagenomic data. This technique, requiring only a single cell, relies on specific organism isolation and whole genome amplification. Single cell genomics has led to numerous genome reconstructions from various ecosystems but is nevertheless not always easily practicable due to contaminations and biased DNA amplification. Finally, gene capture for microbial ecology has also been developed and could contribute to the reconstruction of large DNA fragments or even complete genomes. This review will present promising approaches and results to capture microbial dark matter, an essential step to a better understanding of microbial ecosystem functioning.

Introduction

The evolution has shaped all life on Earth and led to the establishing of the incredible microbial diversity that can be observed nowadays. Microorganisms represent the most diverse and abundant life community. Their small size, as their rapid cellular cycle but especially their metabolic versatility enable them to be widely distributed, colonizing all ecological niches, even the most extreme where they can thrive (Finley, 2002). Their involvement is recognized in various processes, including biogeochemical cycles, trophic network functioning, regulation of populations as pathogenic agents, production of the vast majority of marketed bioactive compounds or else maintenance of global genetic resources.

Consequently, they play a crucial role in organization, evolution and functioning of different ecosystems, and this whatever the nature of these complex environments.

If during many years the exploration of environmental microbial communities has been limited by methodological approaches such as culture-based methods, the 1990s represented a real transition in the field of microbial ecology with the application of molecular biology methods (Pace *et al*., 1995). Indeed, targeting biomarkers of interest such as the small subunit ribosomal RNA (SSU rRNA) genes by fingerprinting (DGGE/TTGE, T-RFLP, ARDRA, ARISA), Fluorescent *In Situ* Hybridization (FISH) or else cloning techniques followed by Sanger sequencing, allowed going into the depth of our knowledge about the structure of microbial populations in various ecosystems (Muyzer *et al*. 1993, Gray et Herwig 1996, Liu *et al*. 1997, Fisher and Triplett 1999, Giraffa and Neviani 2001). Thereafter, the emerging field of metagenomics (Handelsman, 1998) combined with the development of next generation sequencing (NGS) technologies have revolutionized our vision of the prokaryotic diversity (Shokralia *et al*., 2012; Logares *et al*., 2013). Thus, from these, two concepts have emerged. The first one, which appeared following the massive flood of data generated from high throughput sequencing, corresponds to the blossoming of species that are underrepresented in the environment. Indeed, the long tail in rank-abundance curves used to depict diversity highlight that the vast majority of species are present at very low abundance and are thus defined as belonging to the "rare biosphere" (Pedros-Alio, 2006),. . Nevertheless, some studies focused on these rare populations have underlined that a significant part was active in the environment (Campbell *et al*., 2011; Gaidos *et al*., 2011; Hugoni *et al*., 2013), but also that they could become dominant in term of abundance according to the environmental conditions (Teira *et al*., 2007; Hugoni *et al*., 2013). These different observations allowed reconsidering the rare biosphere as not being only a dormant seed bank, but showed that it has an ecological significance (Pedros Alio, 2012) and their roles should so be determined. The second concept, which represents a real question mark for microbial ecologists, is defined as the "microbial dark matter". It corresponds to the discovery of lineages of microorganisms whose functions in environments are still unknown (Rinke *et al*., 2013). Indeed, the accession to the microbial member's identification through a small DNA fragment, such as the 16S rRNA genes, informs in any case on the metabolic potential of these microorganisms and consequently on their functional impact on the environments.

The resolution of the microbial diversity that composes an ecosystem may be possible by means of the advent of the high-throughput DNA sequencing technologies. However, as it was reminded by Lozupone (2012): "Knowing the composition of the microbial community

alone does not necessarily lead to an understanding of its function". Thus, in order to improve the understanding of ecosystem functioning, it is necessary to get deeper insights into the microbial processes and consequently to establish a link between identified microbial populations and metabolic capacities that are present in the environment even at very low abundance levels. To answer to this challenge, complete genome sequencing from microbial isolates has been initiated by projects such as Genomic Encyclopedia of Bacteria and Archaea (Wu *et al*., 2009). Nevertheless, the numerous candidate phyla that compose the major part of microbial diversity are still weakly addressed due to their low abundance in complex environments and/or the incapacity to cultivate them (Rinke *et al*., 2013). Consequently, capturing the microbial dark matter that could be in the rare biosphere and thereafter define functional capabilities by complete genome sequencing represents a real challenge in microbial ecology for better understanding the ecosystem functioning.

Microbial diversity revealed by barcoding

The advent of molecular approaches allowed accessing to populations which were not only restricted to cultured organisms that represent less than 1% of microbial populations (Amann *et al*., 1995; Rappé and Giovannoni, 2003). First culture-independent strategies to characterize microbial communities are based on the study of taxonomic biomarkers such as the SSU rRNA gene. The specific amplification using PCR followed by cloning and sequencing with Sanger method of biomarker was firstly performed to characterize barcodes reflecting the phylogenetic diversity of the studied environment. By this approach, new branches of the life tree called candidate divisions have then been uncovered from various ecosystems. Regarding approach biases, candidate divisions are defined as monophyletic groups that include no cultured organisms, described by at least two nearly complete 16 rRNA gene sequences preferably obtained from two distinct clone libraries and whose name come from the study that firstly identified them (Hugenholtz, 2002; Rappé and Giovannoni, 2003) (Table 1). While some of these phyla still have no cultured representatives, intensive culture efforts have led to the isolation of organisms belonging to candidate divisions revealed by amplicon sequencing such as divisions SAR11 (Rappé et al., 2002), OP5 (Mori et al., 2008) and OP10 (Dunfield et al., 2012). In addition, high-throughput sequencing approaches of targeted biomarker have greatly increased available data to refine classification and especially associate sequences from candidate division with sequences belonging to cultured organism (Dunfield 2012 McDonald 2012). The first application of high-throughput 16S rRNA amplicon sequencing have also revealed the "rare biosphere" from deep-sea samples (Sogin *et*

al., 2006) and the advantages of this technique have rapidly made it the golden standard for screening the microbial diversity. Finally, recent improvements in HTS like the Illumina plateform by merging paired-end reads can generate large sequence (around 500 bp) with higher quality and so increase knowledge about the real microbial communities from complex environment including rare "rare biosphere" and/or dark matter. In spite of these advantages, there are a number of limitations and biases that may be introduced throughout the studies inducing under- or overestimation of the diversity (Hong *et al*., 2009, (Kunin *et al*., 2010) Thus, due to sampling, extraction, amplification conditions, sequencing and analysis biases (Engelbrektson *et al*., 2010; Lee *et al*., 2012, Hazen *et al*, 2013, Wang *et al*., 2013) our vision of the true microbial diversity remains skewed. Other parameters can also impact the representation of the microbial diversity such as the choice of the targeted region (Huse *et al*., 2008; Hamady and Knight, 2009) and the selected primer set (Baker *et al*., 2003). Several 16S rRNA "universal" primers for bacteria or archaea have been proposed and discussed (Baker *et al*., 2003; Sipos *et al*., 2007; Klindworth *et al*., 2013; Mori *et al*., 2013; Wang *et al*., 2014). However, since no universal primer pairs harbor a whole coverage, the best primers must be chosen with the knowledge of biological questions and samples. (Kunin *et al*., 2010).

Bioinformatic tools were therefore developed to overcome these biases (e.g. chimerical sequences, sequence quality, sequencing errors, phylogenetic classifier) (Quince *et al*., 2011; Logares *et al*., 2012), (Edgar *et al*., 2011; Haas *et al*., 2011; Wright *et al*., 2012) (Huse *et al*., 2010; Schloss *et al*., 2011; Edgar, 2013). However, these tools do not ensure a perfect assessment of real microbial diversity (Hazen *et al*., 2013; Poretsky *et al*., 2014).

Diversity revealed by metagenomics

Metagenomic strategies based on direct isolation of nucleic acids from environmental samples and massive sequencing have proven to be powerful tools for exploring ecosystems (Simon and Daniel, 2011). Thus, the unravelling of novel microbial lineages as initially demonstrated by Venter *et al*. (2004) who identified 148 novel phylotypes in the surface water of Sargasso Sea.

The real benefice of metagenomics *versus* 16S rRNA amplicons relies on its ability to bring functional information about microorganism metabolisms and could link microbial community structure to environmental processes (Allen and Banfield, 2005). The power of metagenomics to provide valuable information about uncultivated microbial lineages was

demonstrated by Tyson *et al*. (2004). Applied on a low diversity environment, this landmark study used metagenomics coupled with Sanger sequencing to reconstruct genomes of uncultured bacteria (*i.e. Leptospirillum* group II and *Ferroplasma* group II) revealing complete metabolic pathways and providing insight into their nutritional requirements and biogeochemical functions (Tyson *et al*., 2004; Temperton and Giovannoni, 2012). Other examples as *Candidatus Sulfuricurvum sp.* (Handley *et al*., 2014) and *Thauera* genomes (Mao *et al*., 2014) were reconstructed from more complex environments using second generation sequencing but concern either genomes of abundant population in the studied ecosystem or occurred after enrichment (Supplementary Table X). Indeed, the proportion of assembled reads is directly related to the complexity of the studied microbial community. Thus, for very low complexity environments comprising only few taxa, 85% of the metagenomic reads can be assembled (Tyson *et al*., 2004), whereas for highly complex communities, the assembled proportion of reads generally does not exceed 10%. Even with a tremendous sequencing effort (*i.e*. greater than 12 million reads), unassembled sequences still represent 76% of input reads in complex soil ecosystem (Delmont *et al*., 2012). This weak recovery during metagenomic assembly into large contigs or complete genomes is mainly due to the presence in a same sample of many strains and closely related species. A genome from metagenomic data is therefore always a genome that represents a population and not the genome of a single organism (Sharon and Banfield, 2013).

To overcome these difficulties, innovative assembly strategies can be employed. Binning input reads (*i.e.* gathering reads of similar species) prior to the metagenomic assembly is one of most employed approaches (Kim *et al*., 2013; Nagarajan and Pop, 2013; Segata *et al*., 2013) whereas co-occurrence-based strategies appear to be promising (refs).

Despite constantly improved throughput of the sequencing technologies, the sequencing depth required to provide a comprehensive sample of microbial communities remains inconceivable. It appears therefore necessary to reduce the complexity of metagenomic samples before the DNA extraction, for example by isolation of interest cells such in single-cell genomics.

Single cell genomics

Single Cell Genomics (SCG) emerged as a powerful tool allowing unprecedented access to uncultured microbial communities, either prokaryotic or eukaryotic, and linking the

functions back to the species. To achieve this, single cell approaches involve numerous steps, from the isolation of single cells from the environmental sample to genome sequencing, through cell lysis and whole genome amplification. As described below, each one is critical and needs to be efficiently processed to perform efficient SCG.

Cell isolation

Because it determines the success of later analysis, physical isolation of cells from microbial communities is a major step. Indeed, the different techniques used must enable the efficient and specific selection of cells among all micro-organisms present in the sample. This step can be facilitated by the enrichment of a microbial fraction from the sample which can be carried out through isopycnic density gradient centrifugation for soil samples (Kvist, Ahring, & Lasken, 2007; Podar *et al*., 2007) or tangential filtration for aquatic samples (Giovannoni, DeLong, Schmidt, & Pace, 1990; Martinez-Garcia, Swan, *et al*., 2012). Whatever the chosen isolation technique, cells can be randomly isolated from the environment and further be screened through PCR (Ghai *et al*., 2011; Marcy, Ouverney, *et al*., 2007; Swan *et al*., 2011), or they can be targeted with a variety of fluorescent specific markers given the selection criteria. Thus, cells of interest must be clearly identified among the community based on taxonomic affiliation or special functional capability. The most common approach to achieve this is the use of FISH probes which nevertheless requires the knowledge of specific markers sequences and cannot therefore be applied to all microorganisms (Kalyuzhnaya *et al*., 2006; Kvist *et al*., 2007; Podar *et al*., 2007; Sekar, Fuchs, Amann, & Pernthaler, 2004; Wallner, Fuchs, Spring, Beisker, & Amann, 1997). Other classical approach is the use of fluorescent labelled antibodies and GFP-tagged proteins to screen out communities exclusively regarding metabolic potential (Martinez-Garcia *et al*., 2012). Different techniques permit such isolation of cells and their use mainly depends on the desired throughput and organisms targeted.

Micromanipulation approaches encompass several methods which allow cell selection under continuous high resolution microscopy. As they are easy and cheap, those classical techniques are commonly used in single cell isolation (Shapiro, Biezuner, & Linnarsson, 2013). Micropipetting is the main mechanical micromanipulation method and has shown its high efficiency whatever the type of sample and organism targeted (Grindberg *et al*., 2011; Ishøy, Kvist, Westermann, & Ahring, 2006; Kvist *et al*., 2007; Woyke *et al*., 2010). The other widespread approach, classified in optical micromanipulation methods, is the optical tweezers technology, which uses a focused laser beam for trapping cells (Blainey, Mosier, Potanina, Francis, & Quake, 2011; Huang, Ward, & Whiteley, 2009; Huber, Huber, & Stetter, 2000;

Ishøy *et al*., 2006; Marshall, Blainey, Spormann, & Quake, 2012). The particularity and strength of this technique is that it enables a noncontact manipulation of cells inside sealed vessel preventing contaminations. Those cell isolation techniques, which enable a visual inspection and do not generate mechanical forces, are carried out with the high confidence that cells are well delivered to downstream processing steps and are not damaged (Blainey, 2013; Ishii, *et al.*, 2010). Nevertheless, because cells are isolated one at a time, such micromanipulation approaches are extremely low throughput and very tedious.

Thus, the large majority of single cell isolation approaches relies on Fluorescence Activated Cell Sorting (FACS) that partially addresses disadvantages of micromanipulation devices (Ghai *et al*., 2011; Martínez Martínez, Poulton, Stepanauskas, Sieracki, & Wilson, 2011; Martinez-Garcia, Brazel, Swan, *et al*., 2012; Siegl *et al*., 2011; Swan *et al*., 2011; Woyke *et al*., 2009; Yoon *et al*., 2011). Indeed, this flow cytometry technique has been developed in order to automate the process of single cell selection and therefore significantly increase the isolation throughput (Kalisky *et al*, 2011). Thus, tens of thousands of individual targeted cells can be isolated and delivered within few hours into tubes or microwell plates (Fleming *et al*., 2011; Martinez-Garcia, Swan, *et al*., 2012; Podar *et al*., 2007; Raghunathan *et al*., 2005; Swan *et al*., 2011). This approach has facilitated the isolation without prejudice of all cells from different microbial communities thanks to generic fluorescent markers such as Syto-9 DNA stain (Ghai *et al*., 2011; Martinez-Garcia, Swan, *et al*., 2012; Swan *et al*., 2011). Using FISH probes or other labelled molecules, FACS has also been used to directly select and isolate specific organisms based on their taxonomic affiliation (Kalyuzhnaya *et al*., 2006; Podar *et al*., 2007; Wallner *et al*., 1997) or their particular metabolism (Martinez-Garcia *et al*., 2012). Nevertheless, those targeting approaches highlight the difficulty to avoid nonspecific cell isolation and to reach a high levels of accuracy with FACS machines (Kalyuzhnaya *et al*., 2006; Podar *et al*., 2007). FACS presents other limitations related to the high shear forces induced by the flow which might damage cells and influence the former steps, and the large downstream processing volumes that raise costs and sensitivity to contamination (Blainey, 2013; Yilmaz & Singh, 2012).

Microfluidic devices enable to overcome those last limitation reducing at micro-scale dimensions all processes of cell isolation and providing a sealed environment significantly minimizing the risk of contamination (Agresti *et al*., 2010; Fu, Spence, Scherer, Arnold, & Quake, 1999; Liu *et al*., 2012; Marcy, Ouverney, *et al*., 2007; Yilmaz & Singh, 2012). Over the last few years, these “lab-on-a-chip” approaches have set up new strategies for single cell analysis thanks to the fabrication of microfluidic chips allowing the realization of nanoliter

scale reactions using controlled liquid streams (El-Ali, Sorger, & Jensen, 2006; Shapiro *et al*., 2013). Even efficient and providing inherent advantages, those techniques are only at early stages of development and require further improvements in terms of cell targeting and throughput (Yilmaz & Singh, 2012). The ultimate objective is consequently to optimize a complete and robust microfluidic device that integrates the entire process of single cell genomics analysis (Lecault *et al*, 2012).

Cell lysis

After isolating cells, the next step is to lyse them to access genomic DNA (gDNA). This step is one of the most critical because subsequent genomic analyses depend on the quantity and the quality of gDNA available, especially as prokaryotic cells contain only a few femtograms of gDNA, and that frequently, only a single copy of each genomic locus is present, so any fragmentation and loss impact on later genomic analysis (Blainey, 2013; Yilmaz & Singh, 2012). Conventional bulk cell lysis procedures such as physical disruption (sonication, bead beating, shearing, grinding, high pressure) and chemical handling with ionic surfactants (SDS, Tween) are often irrelevant for single-cell permeation because they subject cells to harsh treatments which cause DNA damages (Blainey, 2013). Thus if DNA is broken in a genomic locus represented only once, genome reconstruction is compromised due to the absence of molecules spanning the break. Freeze/thawing is a moderate disruption method (Mussmann *et al*., 2007) and is often combined with chemical lysis using a strong base like KOH (Kvist *et al*., 2007; Raghunathan *et al*., 2005; Siegl *et al*., 2011). Another approach is enzymatic digestion of cell wall (Fleming *et al*., 2011; Marcy, Ouverney, *et al*., 2007; Swan *et al*., 2011). This is the most gentle lysis method as no mechanical forces are required but its efficiency is consequently low (Stepanauskas, 2012). Further improvements and standardization are needed to ensure a more effective single cell lysis.

Whole genome amplification

A major constraint of single-cell genomic approaches is the absence of any robust single-molecule sequencing technology (Eberwine, Sul, Bartfai, & Kim, 2013). Consequently, because NGS processes require micrograms of DNA as input, direct sequencing of a single microbial genome which typically contains femtograms of DNA is impossible. Single cell Whole Genome Amplification (WGA) is therefore a crucial step before single cell gDNA sequencing. All WGA techniques are based on the use of DNA polymerases with various priming strategies (Blainey, 2013; Van Loo & Voet, 2014). Some of them are PCR-based and

have been successfully applied like Primer Extension Pre-amplification PCR (PEP-PCR) (Dean *et al*., 2002; Hubert, Weber, Schmitt, Zhang, & Arnheim, 1992; Pinard *et al*., 2006; L. I. N. Zhang *et al*., 1992) and Degenerate Oligonucleotide Primed PCR (DOP-PCR) (Dean *et al*., 2002; Klein *et al*., 1999; Pinard *et al*., 2006; Telenius *et al*., 1992) that take advantage of degenerate oligonucleotide primers that obviate the need for ligation reactions or any knowledge of the sequence to be amplified. Nevertheless, single microbial cell WGA is preferentially performed by Multiple Displacement Amplification (MDA) (Blainey *et al*., 2011; Martínez Martínez *et al*., 2011; Mason *et al*., 2012; Rinke *et al*., 2014; Rodrigue *et al*., 2009). This isothermal amplification method uses random primers and the Phi29 DNA polymerase derived from the *Bacillus subtilis* bacteriophage which has a strong strand displacement activity (Dean, Nelson, Giesler, & Lasken, 2001; Yilmaz & Singh, 2012). This method generates long overlapping amplicons averaging 12 kb (Dean *et al*., 2002) and going up to 70 kb in length (Blanco *et al*, 1989) with high fidelity thanks to the 3'-5' exonuclease proofreading activity of the Phi29 DNA polymerase (L Blanco & Salas, 1984). Such high molecular weight amplified DNA molecules and the important yields obtained are ideal for library construction (Pinard *et al*., 2006). Although it is clearly the most favoured method for single cell WGA, MDA is not completely unbiased (Pinard *et al*., 2006). The key drawbacks mainly include uneven genome coverage, chimeric sequences, and contamination issues (Lasken & Stockwell, 2007; Marcy, Ishoey, *et al*., 2007; Raghunathan *et al*., 2005; Woyke *et al*., 2009).

The first bias of MDA and all WGA methods is related to the sample contamination itself, that however can be significantly reduced by shrinking the sample volumes to the nanoliter scale at each step, from cell isolation to genome amplification (Blainey & Quake, 2011; Marcy, Ouverney, *et al*., 2007). Minimizing the reaction volumes concentrates the single cell or single genome to be amplified and consequently reduces the proportion of contaminates in the sample. Such approach applied in microfluidic devices has shown its efficiency with low sample contamination detected (Marcy, Ishoey, *et al*., 2007). Nevertheless, given the high concentration of contaminating DNA fragments in commercial reagents, varying from 5 to 50 fragments per microliter, a simple volume reduction does not systematically eliminates all contaminates (Blainey & Quake, 2011). Even dramatically decreasing volumes, reagent contaminating DNA can represent a significant fraction likely to be detrimental to subsequent genome assembly, and representing up to the entirety of DNA products in case of lysis failure (Blainey, 2013). Therefore, a usual practice to limit such contamination consists in a UV-pretreatment which is applied to reagents, but also to tubes,

plates and all small common equipment to get rid of other sources of contaminating DNA. Contaminations from laboratory environment and instruments can be effectively addressed thanks to a cleanup with bleach of any device used, such as dedicated pipetters or isolating devices, followed by a rinsing with UV-treated water to efficiently remove potential DNA contamination (Rodrigue *et al*., 2009; Woyke *et al*., 2011; Motley *et al*., 2014). Despite taking extensive precautions, DNA contamination remains a key problem in single cell approaches and must be kept in mind at every single step. The last resort is computational identification and elimination of such sequences (Podar *et al*., n.d.; Woyke *et al*., 2010).

MDA reactions are also subjected to stochastic and sequence dependent priming at the early stages of the reaction, resulting in substantial variations in the amplifications of the different regions of the Single Amplified Genome (SAG) (Pinard *et al*., 2006; Podar, Keller, & Hugenholtz, n.d.; Rodrigue *et al*., 2009). Several strategies have been proposed to minimize this bias and even out amplification. The first one is to decrease MDA reaction volumes to increase template concentration, and consequently improve genomic coverage (Pinard *et al*., 2006). Its efficiency has been particularly well described with the use of microfluidic devices which by reducing reactions volumes to the nanoliter scale have significantly improved MDA coverage (Marcy, Ishoey, *et al*., 2007; Marcy, Ouverney, *et al*., 2007). Supplementing those reactions with crowding agents like trehalose or polyethylene glycol (Pan *et al*., 2008) can provide even more homogenous amplification. Another efficient strategy to reduce uneven representation due to single cell WGA is post-amplification normalization to avoid the disproportionately large effort directed towards sequencing a small fraction of the gDNA (Rodrigue *et al*., 2009; Swan *et al*., 2011). It is carried out thanks to the activity of a duplex-specific nuclease which removes highly abundant double stranded DNA. This approach can be completed by an *in silico* normalization consisting in a removal of overrepresented reads (Rodrigue *et al*., 2009; Swan *et al*., 2011) and by the use of dedicated *de novo* single cell genome assembly software which facilitate assembly of SAGs (Harrington, Arumugam, Raes, Bork, & Relman, 2010). The presence of chimeras, which correspond to the second bias of the MDA, can also be reduced by up to 80% by post-MDA treatments such as the use of S1 nuclease that should cleave the single stranded region that connects the two segments of the chimera (K. Zhang *et al*., 2006) (Lasken & Stockwell, 2007). Finally, to overcome both uneven genome coverage and chimeric sequences, the pooling of few cells or MDA amplified genomes from different individuals of the same species or clonal populations has also been suggested. Increasing copy number provides a better representation of each locus resulting in an enhanced coverage and consequently an easier computational elimination of chimeras

(Podar *et al*.; Tyson *et al*., 2004; Raghunathan *et al*., 2005; Warnecke and Hugenholtz, 2007; Woyke *et al*., 2010; Morales and Holben, 2011). Despite such cells are expected to have only slight genome variations, the resulting assembly must be considered as a composite genome given the important cellular heterogeneities in natural populations, particularly for bacterial species whose genome have been proven to be exceedingly variable (Eberwine *et al*., 2013; Marcy, Ouverney, *et al*., 2007).

Sequencing and assembly

Among the multiple sequencing technologies, all are suited for SCG and the choice between all them must be done considering the desired use of data (Stepanauskas, 2012). Nevertheless, most of the recent software developed for single cell whole genome assembly are preferentially dedicated to paired-end Illumina reads because of the ability to generate high-quality draft assemblies with this approach (Bankevich *et al*., 2012; Chitsaz *et al*., 2013; Peng *et al* , 2012; Swan *et al*., 2011). Those bioinformatics tools enable data treatment, from contamination, uneven amplification bias and chimera suppression, to assembly and annotation, and demonstrate a higher performance in *de novo* assembly than usual assembly software (Blainey, 2013). Although all sequences obtained from single cell approaches come with certainty from a unique cell unlike for metagenomics, the success of genome recovery varies widely, from 12% to a complete genome (Eloe *et al*., 2011; Hongoh *et al*., 2008; Woyke *et al*., 2010; Youssef, Blainey, Quake, & Elshahed, 2011). It is directly linked to the correct achievement of each one of the different steps leading to a SAG and to the specific properties of cells, such as cell wall structure and polyploidy (Woyke *et al*., 2010). Genome reconstruction enhancement can be achieved through gap closure (Hongoh *et al*., 2008; Woyke *et al*., 2010) or by a combination of approaches including use of multiple cells or metagenomic data sets.

Microbial ecology contribution

Thanks to its ability to reconstruct partial or even complete genome from only one targeted organism without the complications of untangling data from multiple cells, SCG have been proved to be a powerful tool to explore many fields of microbial ecology. This approach has broadly been used to provide gDNA sequences and some nearly complete genomes for numerous major taxa with no sequenced representative genomes in databases. One interesting example is the first genome reconstruction of five low salinity type Ammonia-Oxidizing Archaea (AOA), named Candidatus *Nitrosoarchaeum limnia* SFB1, from an enrichment

culture (Blainey *et al*., 2011). Each WGA led to a 60% genome recovery, and combination of the five genomic sequence data permitted to rebuild a high quality draft genome assembly representing more than 95% of the complete sequence. Nevertheless, such studies concerning cultivated organisms remain scarce because bypassing the conventional cultivation step, SCG presents a great advantage which is the access to the genetic makeup of uncultivated microbes directly isolated from environmental samples. For instance, McLean *et al*. (2013) assembled the first partial genome from a member of the globally distributed candidate phylum TM6, and Woyke *et al*. (2009) the one of coastal ocean waters *Flavobacteria*. Others studies provided genomes of filamentous *Beggiatoa* (Mussmann *et al*., 2007), candidate phylum *Poribacteria* (Siegl *et al*., 2011) or even SAR342 clade of *Deltaproteobacteria* (Chitsaz *et al*., 2013). Single cell approaches are particularly interesting for their capability to target microbial dark matter and particularly microorganisms that could belong to the rare biosphere as it was underlined by Marcy *et al*. (2007) and Podar *et al*. (2007). Indeed, they demonstrated the possibility to isolate from different environments and through different techniques (*i.e*. laser tweezers and microfluidics) uncultured TM7 cells, present between 0.7 and 1.9% in the ecosystems studied. They enabled the partial assembly of those two genomes, what had never been done before.

The even partial reconstruction of genomes provides essential information on the micro-organisms metabolic capacities and consequently on their role in the environment. A typical example is the partial genome assembly of a single cell belonging to the yet uncultured candidate division OP11 widely distributed in terrestrial and marine ecosystems (Youssef *et al*., 2011). If few information were available regarding those poorly understood bacteria, their various metabolic capabilities allowing stress response, multiple antibiotic production and resistance mechanisms enhancing their survival in diverse and competitive habitat have thus been shown. The genomic determination of the heterotrophy of the novel marine protist group of Picobiliphyta, originally defined as phototrophic based on its ultrastructure description has also been highlighted (Yoon *et al*., 2011). Broadly, SCG enables to collect information on the probable contribution of targeted organisms within ecosystems and their impact on trophic networks. It has well been illustrated by Swan *et al*. (2011), who evidenced chemolithoautotrophy pathways in uncultured *Deltaproteobacteria* cluster SAR324, *Gammaproteobacteria* clusters ARCTIC96BD-19, Agg47, and some *Oceanospirillales* that constitute a major fraction of dark ocean's biomass and could significantly contribute to carbon cycling in the ocean. Also, all of the ubiquitous freshwater bacterioplankton clusters such as *Actinobacteria* acI, *Polynucleobacter spp*. and LD12 contain at least a

photoheterotroph membersuggesting that photoheterotrophy is widespread in the euphotic zone of temperate freshwater lakes (Martinez-Garcia *et al*. 2012). Thus, such precise knowledge of microbial metabolic capabilities through single cell approaches open up new horizons for development of biotechnological applications like conversion of cellulosic and polysaccharidic biomass into biofuels (Hess *et al*., 2011; Martinez-Garcia, Brazel, *et al*., 2012), and potentially for guided cultivation of uncultured microorganisms.

SCG also permits to better understand the interactions between organisms such as those occurring in symbiosis (commensalism, mutualism and parasitism). Indeed, the simultaneous access to genomes belonging to host and symbionts which are usually not accessible through cultivation independent techniques, provides unique insights into their relations. Thus, Martinez-Garcia, *et al*. (2012) revealed single cell scale interactions between diverse heterotrophic marine eukaryotes and bacteria such as the association between *Pelagibacter ubique* and MAST-4 protist or *Actinobacterium* and chrysophytes. Similarly, Hongoh *et al*. (2008) assembled complete genomes of intracellular symbionts of protist, which themselves are termite gut symbionts. An interesting approach to discriminate and isolate *Emiliania huxleyi* microalga cells infected by viruses and to analyze the viral genome allowing host–virus interactions to be studied was also developed (Martínez Martínez *et al*. (2011).

Such interactions can induce genome rearrangements in micro-organism populations which rapidly diversify, like gene insertion through horizontal gene transfer (HGT) and loss. SCG is particularly well suited to detect that kind of genomic variability in natural microbial populations compared with metagenomics due to its capability to resolve fine-scale heterogeneity at the population level without ambiguity. A good illustration of this resolution is the first detection of HGT and recombination of rhodopsin genes in freshwater *Gammaproteobacteria*, *Betaproteobacteria* and *Actinobacteria* (Martinez-Garcia, Swan, *et al*., 2012). Even more precisely, single cell approaches revealed single nucleotide polymorphisms in ammonia oxidizing *Archaea* (Blainey *et al*., 2011), segmented filamentous bacteria (Pamp, Harrington, Quake, Relman, & Blainey, 2012) and the *Candidatus Sulcia muelleri* DMIN symbiont (Woyke *et al*., 2010) at the population level, which combined with genomic data can confidently resolve phylogenies and inform about organisms evolution. Those examples illustrate the power of SCG in providing information concerning fine-scale structure of microbial communities thanks to its ability to generate high quality non-composite reference genomes.

Because SCG possesses strengths and metagenomics others, the integration of the two methods provides interesting insights into microbial diversity and makes up for defaults of each one. The first advantage of this complementarity is the possibility to combine genomic data provided by each method to reach complete genome assemblies. Metagenomic reads can be incorporated into single-cell data as for Candidatus *Nitrosoarchaeum limnia* SFB1 (Blainey *et al*., 2011), and in other cases, single cell data can be in turn used to guide and validate genome reconstruction from metagenomic data, like for SAR86 genome (Dupont *et al*., 2012; Hess *et al*., 2011). The second one is the possibility to use SCG as a specific tool to target organisms evidenced in metagenomic analysis and thus explain communities functioning. For instance, Mason *et al*. (2012) used single cell approaches to specifically characterize and confirm the hydrocarbon degradation potential of the dominant Oceanospirillales identified with metagenomics during the Deepwater Horizon oil spill. Likewise, thanks to metagenomics, Ghai *et al*. (2011) discovered a new abundant micro-organisms in hypersaline saltern ponds whose photoheterotrophic and polysaccharide-degrading lifestyle has been investigated with SCG.

SGC thus appears as a particularly interesting and promising field for understanding the ecosystem functioning, with nevertheless remaining challenges. This technique still requires further technological advances to overcome technological limitations making it easily applicable, and to efficiently remove all bias complicating genome reconstruction. Linking specific microbial populations to environmental processes does not necessary imply billions of random-shotgun metagenomics reads or SCG expertise. Gene capture undeniably represents a promising genomic-scale sequence enrichment strategy that could contribute to the reconstruction of large DNA fragments or even complete genomes from complex environments.

Gene capture as an innovative approach for capturing the microbial dark matter

Gene capture approaches, which were firstly developed for resequencing application, rely on either solid phase (Albert *et al*., 2007; Okou *et al*., 2007; Mokry *et al*., 2010) or solution phase hybridization (Tewhey *et al*., 2009; Gnirke *et al*., 2009) of nucleic acid capture probes to the targeted DNA sequences. Denonfoux *et al*. (2013) described the first adaptation of a gene capture method for the selective enrichment of a target-specific genomic locus from a complex environmental metagenomic DNA (Denonfoux *et al*., 2013). Thus, a Solution

Hybrid Selection (SHS) method was applied to a lacustrine environment targeting the methyl coenzyme M reductase subunit A (*mcrA)* gene revealing higher methanogen community diversity than previously observed with other methods. The enrichment performance (> 41%) demonstrated the relevance of such method compared to a random-shotgun metagenomic approach (0.003%). Applied to other genomic loci, including phylogenetic markers such as 16S or 18S rRNA genes, enrichment performance of the gene capture can be greater than 90% (personal data).

The success of a gene capture experiment in microbial ecology strongly depends on the high-quality probe set encompassing variant specific and explorative probes as it was shown by Denonfoux *et al*. (2013). Contrary to the resequencing of regions from complex eukaryotic genomes where tiling design strategies are sufficient, environmental capture probes must combine both sensitivity (*i.e.* probes should detect low abundance targets in complex mixtures) and specificity (*i.e.* probes should not cross-hybridize with non-target sequences) (Parisot *et al*., 2014). Moreover, taking advantage of the exponential growth of sequencing data, explorative probe design strategies for gene capture offer the opportunity to survey both known and unknown sequences (Dugat-Bony *et al*., 2012). Such strategies use the sequence variability within the targeted sequences to define new combinations potentially present in natural environments and that have not yet been described and deposited in public databases. Two main probe design software allow defining oligonucleotides harbouring these criteria: KASpOD (Parisot *et al*., 2012) and HiSpOD (Dugat-Bony *et al*., 2011).

Adaptation of the gene capture for microbial ecology allowed a better taxonomic and functional description of the studied microbial community including the rare biosphere and unknown sequences (Denonfoux *et al*., 2013). Additionally, gene capture allows recovering large DNA fragments highlighting the ability to extend beyond the initial targeted biomarker gene sequence and thus facilitate the discovery of new genes or genomic organizations. Initially coupled with pyrosequencing, the emergence of third generation sequencing platforms and the possibility to sequence longer DNA sequences without library construction (McCarthy, 2010; Schadt *et al*., 2010; Morey *et al*., 2013) should provide real benefits to the gene capture for recovering large genomic regions or even complete microbial genomes allowing then to access to the metabolic capacities of microorganisms and consequently to better understand their implications in the ecosystems

Concluding Remarks

The drastically decreasing costs of the current ultra-high throughput sequencing technologies greatly improved the resolution for microbial diversity description. Amplicon sequencing of functional or phylogenetic biomarkers gives some insights into the microbial community diversity but it lacks information to relate this diversity to environmental processes. Metagenomics can overcome this limitation but hundreds or thousands of sequencing runs are still necessary to provide a comprehensive exploration of complex ecosystems involving handling billions of nucleic acid sequences fragments and their bioinformatic bottlenecks. Based on the results obtained from amplicon or shotgun sequencing, more targeted approaches including single-cell genomics or gene capture can be carried out to focus on particular populations. Such promising strategies contributing to the reconstruction of large genomic fragments or even complete genomes provide valuable information about uncultivated microbial lineages.

Evolution of the techniques has enabled unprecedented access to microbial communities and has allowed scientists to ask questions previously thought impossible to answer. The rare biosphere now becomes visible and complete genomes have been reconstructed principally for dominant micro-organisms in explored ecosystems. The ability of these different techniques (metagenomics, single cell or gene capture) to link ecosystem processes to individual microbial populations allow the ecological studies to illuminate the microbial dark matter and subsequently to better understand the ecosystem functioning. However, every technique discussed in this review has strengths and weaknesses that must be taken into account. Until the development of an innovative technology allowing to unambiguously elucidate both community structure and function, the best experimental design must be chosen with the knowledge of biological questions and samples.

References

Allen, E.E. and Banfield, J.F. (2005) Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.* **3**: 489–98.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**: 143–69.

Baker, G.C., Smith, J.J., and Cowan, D.A. (2003) Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* **55**: 541–555.

Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al*. (2000) Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science (80-.).* **289**: 1902–1906.

Brakenhoff, R.H., Schoenmakers, J.G., and Lubsen, N.H. (1991) Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res.* **19**: 1949.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al*. (2010) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108**: 4516–4522.

Cline, J., Braman, J.C., and Hogrefe, H.H. (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24**: 3546–51.

Delmont, T.O., Prestat, E., Keegan, K.P., Faubladier, M., Robe, P., Clark, I.M., *et al*. (2012) Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J.* **6**: 1677–87.

Dojka, M. a, Hugenholtz, P., Haack, S.K., and Pace, N.R. (1998) Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64**: 3869–77.

Dunfield, P.F., Tamas, I., Lee, K.C., Morgan, X.C., McDonald, I.R., and Stott, M.B. (2012) Electing a candidate: a speculative history of the bacterial phylum OP10. *Environ. Microbiol.* **14**: 3069–80.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–200.

Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H., and Hugenholtz, P. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* **4**: 642–7.

Fieseler, L., Quaiser, A., Schleper, C., and Hentschel, U. (2006) Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum Poribacteria by environmental genomics. *Environ. Microbiol.* **8**: 612–24.

Fuhrman, J.E.D.A., Mccallum, K., and Davis, A.A. (1993) Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans . *Appl. Environ. Microbiol.* **59**: 1294–1302.

Ghai, R., Pašić, L., Fernández, A.B., Martin-Cuadrado, A.-B., Mizuno, C.M., McMahon, K.D., *et al*. (2011) New abundant microbial groups in aquatic hypersaline environments. *Sci. Rep.* **1**: 135.

Hamady, M. and Knight, R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19**: 1141–52.

Handley, K.M., Bartels, D., O'Loughlin, E.J., Williams, K.H., Trimble, W.L., Skinner, K., *et al*. (2014) The complete genome sequence for putative H2 - and S-oxidizer Candidatus Sulfuricurvum sp., assembled de novo from an aquifer-derived metagenome. *Environ. Microbiol.*

Harris, J.K., Kelley, S.T., and Pace, N.R. (2004) New Perspective on Uncultured Bacterial Phylogenetic Division OP11. *Appl. Environ. Microbiol.* **70**: 845–849.

Hugenholtz, P. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* **53**: 289–293.

Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**: REVIEWS0003.

Hugenholtz, P., Pitulle, C., Hershberger, K.L., and Pace, N.R. (1998) Novel Division Level Bacterial Diversity in a Yellowstone Hot Spring. *J. Bacteriol.* **180**: 366–376.

Huse, S.M., Dethlefsen, L., Huber, J. a, Mark Welch, D., Welch, D.M., Relman, D.A., and Sogin, M.L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4**: e1000255.

Janssen, P.H., Yates, P.S., Grinton, B.E., Taylor, P.M., and Sait, M. (2002) Improved Culturability of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions Acidobacteria , Actinobacteria , Proteobacteria , and Verrucomicrobia. *Appl. Environ. Microbiol.* **68**: 2391–2396.

Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002) Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. *Science* **296**: 1127–9.

Kim, M., Lee, K.-H., Yoon, S.-W., Kim, B.-S., Chun, J., and Yi, H. (2013) Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform.* **11**: 102–13.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**: e1.

Lee, C.K., Herbold, C.W., Polson, S.W., Wommack, K.E., Williamson, S.J., McDonald, I.R., and Cary, S.C. (2012) Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* **7**: e44224.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H., *et al*. (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.*

Mao, Y., Xia, Y., Wang, Z., and Zhang, T. (2014) Reconstructing a Thauera genome from a hydrogenotrophic-denitrifying consortium using metagenomic sequence data. *Appl. Microbiol. Biotechnol.*

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., *et al*. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**: 610–8.

Mori, H., Maruyama, F., Kato, H., Toyoda, A., Dozono, A., Ohtsubo, Y., *et al*. (2013) Design and Experimental Application of a Novel Non-Degenerate Universal Primer Set that Amplifies Prokaryotic 16S rRNA Genes with a Low Possibility to Amplify Eukaryotic rRNA Genes. *DNA Res.* **21**: 217–27.

Mori, K., Sunamura, M., Yanagawa, K., Ishibashi, J., Miyoshi, Y., Iino, T., *et al*. (2008) First cultivation and ecological investigation of a bacterium affiliated with the candidate phylum OP5 from hot springs. *Appl. Environ. Microbiol.* **74**: 6223–9.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified. *Nat. Rev. Genet.* **14**: 157–67.

Ohkuma, M. and Kudo, T. (1996) Phylogenetic Diversity of the Intestinal Bacterial Community in the Termite Reticulitermes speratus. *Appl. Environ. Microbiol.* **62**: 461–468.

Poretsky, R., Rodriguez-R, L.M., Luo, C., Tsementzi, D., and Konstantinidis, K.T. (2014) Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS One* **9**: e93827.

Qiu, X., Wu, L., Huang, H., Mcdonel, P.E., Palumbo, A. V, Tiedje, J.M., and Zhou, J. (2001) Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning. *Appl. Environ. Microbiol.* **67**: 880–887.

Rappé, M.S., Connon, S. a, Vergin, K.L., and Giovannoni, S.J. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630–3.

Rappé, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**: 369–94.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**: 525–52.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., and Chisholm, S.W. (2010) Unlocking short read sequencing for metagenomics. *PLoS One* **5**: e11840.

Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., and Huttenhower, C. (2013) Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**: 666.

Sharon, I. and Banfield, J.F. (2013) Microbiology. Genomes from metagenomics. *Science* **342**: 1057–8.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* **60**: 341–50.

Sogin, M.L., Morrison, H.G., Huber, J. a, Mark Welch, D., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* **103**: 12115–20.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J. a, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wang, Y., Tian, R.M., Gao, Z.M., Bougouffa, S., and Qian, P.-Y. (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PLoS One* **9**: e90053.

Wright, E.S., Yilmaz, L.S., and Noguera, D.R. (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl. Environ. Microbiol.* **78**: 717–25.

Xu, J. (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol. Ecol.* **15**: 1713–31.

Agresti, J. J., Antipov, E., Abate, A. R., Ahn, K., Rowat, A. C., Baret, C., … Aga, G. A. L. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, *107*(14), 6550–6550. doi:10.1073/pnas.1002891107

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a, Dvorkin, M., Kulikov, A. S., … Pevzner, P. a. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *19*(5), 455–77. doi:10.1089/cmb.2012.0021

Blainey, P. C. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, *37*(3), 407–27. doi:10.1111/1574-6976.12015

Blainey, P. C., Mosier, A. C., Potanina, A., Francis, C. A., & Quake, S. R. (2011). Genome of a Low-Salinity Ammonia-Oxidizing Archaeon Determined by Single-Cell and Metagenomic Analysis, *6*(2). doi:10.1371/journal.pone.0016626

Blainey, P. C., & Quake, S. R. (2011). Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Research*, *39*(4), e19. doi:10.1093/nar/gkq1074

Blanco, L., Bernads, A., Lharo, J. M., Martins, G., & Garmendia, C. (1989). Highly Efficient DNA Synthesis by the Phage 429 DNA Polymerase.

Blanco, L., & Salas, M. (1984). Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(17), 5325–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=391696&tool=pmcentrez&rendertype=abstract

Chitsaz, H., Yee-greenbaum, J. L., Tesler, G., Lombardo, M., Dupont, C. L., Badger, J. H., … Lasken, R. S. (2013). De novo assembly of bacterial genomes from single cells. *National Institutes of Health*, *29*(10), 915–921. doi:10.1038/nbt.1966.De

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, a F., Bray-Ward, P., … Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(8), 5261–6. doi:10.1073/pnas.082089499

Dean, F. B., Nelson, J. R., Giesler, T. L., & Lasken, R. S. (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification, 1095–1099. doi:10.1101/gr.180501.4

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Richter, R. A., Valas, R., … Venter, J. C. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal*, *6*(6), 1186–99. doi:10.1038/ismej.2011.189

Eberwine, J., Sul, J.-Y., Bartfai, T., & Kim, J. (2013). The promise of single-cell sequencing. *Nature Methods*, *11*(1), 25–27. doi:10.1038/nmeth.2769

El-Ali, J., Sorger, P. K., & Jensen, K. F. (2006). Cells on chips. *Nature*, *442*(7101), 403–11. doi:10.1038/nature05063

Eloe, E. a, Fadrosh, D. W., Novotny, M., Zeigler Allen, L., Kim, M., Lombardo, M.-J., … Bartlett, D. H. (2011). Going deeper: metagenome of a hadopelagic microbial community. *PloS One*, *6*(5), e20388. doi:10.1371/journal.pone.0020388

Fleming, E. J., Langdon, A. E., Martinez-Garcia, M., Stepanauskas, R., Poulton, N. J., Masland, E. D. P., & Emerson, D. (2011). What's new is old: resolving the identity of Leptothrix ochracea using single cell genomics, pyrosequencing and FISH. *PloS One*, *6*(3), e17769. doi:10.1371/journal.pone.0017769

Fu, a Y., Spence, C., Scherer, a, Arnold, F. H., & Quake, S. R. (1999). A microfabricated fluorescence-activated cell sorter. *Nature Biotechnology*, *17*(11), 1109–11. doi:10.1038/15095

Garcia, S. L., McMahon, K. D., Martinez-Garcia, M., Srivastava, A., Sczyrba, A., Stepanauskas, R., … Warnecke, F. (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *The ISME Journal*, *7*(1), 137–47. doi:10.1038/ismej.2012.86

Ghai, R., Pašić, L., Fernández, A. B., Martin-Cuadrado, A.-B., Mizuno, C. M., McMahon, K. D., … Rodríguez-Valera, F. (2011). New abundant microbial groups in aquatic hypersaline environments. *Scientific Reports*, *1*, 135. doi:10.1038/srep00135

Giovannoni, S. J., DeLong, E. F., Schmidt, T. M., & Pace, N. R. (1990). Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Applied and Environmental Microbiology*, *56*(8), 2572–5. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=184769&tool=pmcentrez&rendertype=abstract

Grindberg, R. V, Ishoey, T., Brinza, D., Esquenazi, E., Coates, R. C., Liu, W., … Gerwick, W. H. (2011). Single cell genome amplification accelerates identification of the

apratoxin biosynthetic pathway from a complex microbial assemblage. *PloS One*, *6*(4), e18565. doi:10.1371/journal.pone.0018565

Harrington, E. D., Arumugam, M., Raes, J., Bork, P., & Relman, D. a. (2010). SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics (Oxford, England)*, *26*(23), 2979–80. doi:10.1093/bioinformatics/btq564

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., … Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N.Y.)*, *331*(6016), 463–7. doi:10.1126/science.1200387

Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Taylor, T. D., Kudo, T., … Ohkuma, M. (2008). Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(14), 5555–60. doi:10.1073/pnas.0801389105

Huang, W. E., Ward, A. D., & Whiteley, A. S. (2009). Raman tweezers sorting of single microbial cells. *Environmental Microbiology Reports*, *1*(1), 44–9. doi:10.1111/j.1758-2229.2008.00002.x

Huber, R., Huber, H., & Stetter, K. O. (2000). Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiology Reviews*, *24*(5), 615–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11077154

Hubert, R., Weber, J. L., Schmitt, K., Zhang, L., & Arnheim, N. (1992). A new source of polymorphic DNA markers for sperm typing: analysis of microsatellite repeats in single cells. *American Journal of Human Genetics*, *51*(5), 985–91. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1682826&tool=pmcentrez&rendertype=abstract

Ishii, S., Tago, K., & Senoo, K. (2010). Single-cell analysis and isolation for microbiology and biotechnology: methods and applications. *Applied Microbiology and Biotechnology*, *86*(5), 1281–92. doi:10.1007/s00253-010-2524-4

Ishøy, T., Kvist, T., Westermann, P., & Ahring, B. K. (2006). An improved method for single cell isolation of prokaryotes from meso-, thermo- and hyperthermophilic environments using micromanipulation. *Applied Microbiology and Biotechnology*, *69*(5), 510–4. doi:10.1007/s00253-005-0014-x

Kalyuzhnaya, M. G., Zabinsky, R., Bowerman, S., Baker, D. R., Lidstrom, M. E., & Chistoserdova, L. (2006). Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations. *Applied and Environmental Microbiology*, *72*(6), 4293–301. doi:10.1128/AEM.00161-06

Klein, C. a, Schmidt-Kittler, O., Schardt, J. a, Pantel, K., Speicher, M. R., & Riethmüller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(8), 4494–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16360&tool=pmcentrez&rendertype=abstract

Kvist, T., Ahring, B. K., & Lasken, R. S. (2007). Specific single-cell isolation and genomic amplification of uncultured microorganisms, 926–935. doi:10.1007/s00253-006-0725-7

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology*, *7*, 19. doi:10.1186/1472-6750-7-19

Lecault, V., White, A. K., Singhal, A., & Hansen, C. L. (2012). Microfluidic single cell analysis: from promise to practice. *Current Opinion in Chemical Biology*, *16*(3-4), 381–90. doi:10.1016/j.cbpa.2012.03.022

Liu, P., Meagher, R. J., Light, Y. K., Yilmaz, S., Chakraborty, R., Arkin, A. P., … Singh, A. K. (2012). Microfluidic fluorescence in situ hybridization and flow cytometry (μFlowFISH), *11*(16), 2673–2679. doi:10.1039/c1lc20151d.Microfluidic

Marcy, Y., Ishoey, T., Lasken, R. S., Stockwell, T. B., Walenz, B. P., Halpern, A. L., … Quake, S. R. (2007). Nanoliter Reactors Improve Multiple Displacement Amplification of Genomes from Single Cells, *3*(9). doi:10.1371/journal.pgen.0030155

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., … Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11889–94. doi:10.1073/pnas.0704662104

Marshall, I. P. G., Blainey, P. C., Spormann, A. M., & Quake, S. R. (2012). A Single-cell genome for Thiovulum sp. *Applied and Environmental Microbiology*, *78*(24), 8555–63. doi:10.1128/AEM.02314-12

Martínez Martínez, J., Poulton, N. J., Stepanauskas, R., Sieracki, M. E., & Wilson, W. H. (2011). Targeted sorting of single virus-infected cells of the coccolithophore Emiliania huxleyi. *PloS One*, *6*(7), e22520. doi:10.1371/journal.pone.0022520

Martinez-Garcia, M., Brazel, D. M., Swan, B. K., Arnosti, C., Chain, P. S. G., Reitenga, K. G., … Stepanauskas, R. (2012). Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. *PloS One*, *7*(4), e35314. doi:10.1371/journal.pone.0035314

Martinez-Garcia, M., Brazel, D., Poulton, N. J., Swan, B. K., Gomez, M. L., Masland, D., … Stepanauskas, R. (2012, March). Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *The ISME Journal*. doi:10.1038/ismej.2011.126

Martinez-Garcia, M., Swan, B. K., Poulton, N. J., Gomez, M. L., Masland, D., Sieracki, M. E., & Stepanauskas, R. (2012). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal*, *6*(1), 113–23. doi:10.1038/ismej.2011.84

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S. G., Dubinsky, E. a, Fortney, J. L., … Jansson, J. K. (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *The ISME Journal*, *6*(9), 1715–27. doi:10.1038/ismej.2012.59

Morales, S. E., & Holben, W. E. (2011). Linking bacterial identities and ecosystem processes: can "omic" analyses be more than the sum of their parts? *FEMS Microbiology Ecology*, *75*(1), 2–16. doi:10.1111/j.1574-6941.2010.00938.x

Mussmann, M., Hu, F. Z., Richter, M., de Beer, D., Preisler, A., Jørgensen, B. B., … Ehrlich, G. D. (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biology*, *5*(9), e230. doi:10.1371/journal.pbio.0050230

Pamp, S. J., Harrington, E. D., Quake, S. R., Relman, D. a, & Blainey, P. C. (2012). Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome Research*, *22*(6), 1107–19. doi:10.1101/gr.131482.111

Pan, X., Urban, A. E., Palejev, D., Schulz, V., Grubert, F., Hu, Y., … Weissman, S. M. (2008). A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(40), 15499–504. doi:10.1073/pnas.0808028105

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, *28*(11), 1420–8. doi:10.1093/bioinformatics/bts174

Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., Tartaro, K. R., Plant, R. N., … Leamon, J. H. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, *7*, 216. doi:10.1186/1471-2164-7-216

Podar, M., Abulencia, C. B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J. a, … Keller, M. (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology*, *73*(10), 3205–14. doi:10.1128/AEM.02985-06

Podar, M., Keller, M., & Hugenholtz, P. (n.d.). Single Cell Whole Genome Amplification of Uncultivated Organisms. doi:10.1007/7171

Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M., & Lasken, R. S. (2005). Genomic DNA Amplification from a Single Bacterium, *71*(6), 3342–3347. doi:10.1128/AEM.71.6.3342

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., … Woyke, T. (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nature Protocols*, *9*(5), 1038–48. doi:10.1038/nprot.2014.067

Rodrigue, S., Malmstrom, R. R., Berlin, A. M., Birren, B. W., Henn, M. R., & Chisholm, S. W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PloS One*, *4*(9), e6864. doi:10.1371/journal.pone.0006864

Sekar, R., Fuchs, B. M., Amann, R., & Pernthaler, J. (2004). Flow Sorting of Marine Bacterioplankton after Fluorescence In Situ Hybridization, *70*(10), 6210–6219. doi:10.1128/AEM.70.10.6210

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews. Genetics*, *14*(9), 618–30. doi:10.1038/nrg3542

Siegl, A., Kamke, J., Hochmuth, T., Piel, J., Richter, M., Liang, C., … Hentschel, U. (2011). Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *The ISME Journal*, *5*(1), 61–70. doi:10.1038/ismej.2010.95

Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology*, *15*(5), 613–20. doi:10.1016/j.mib.2012.09.001

Swan, B. K., Martinez-Garcia, M., Preston, C. M., Sczyrba, a., Woyke, T., Lamy, D., … Stepanauskas, R. (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science*, *333*(6047), 1296–1300. doi:10.1126/science.1203690

Tamminen, M., & Virta, M. (n.d.). Single gene-based distinction of individual microbial genomes from a mixed population of microbial cells, 1–17.

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. a, & Tunnacliffe, a. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, *13*(3), 718–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1639399

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., … Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. doi:10.1038/nature02340

Van Loo, P., & Voet, T. (2014). Single cell analysis of cancer genomes. *Current Opinion in Genetics & Development*, *24C*, 82–91. doi:10.1016/j.gde.2013.12.004

Wallner, G., Fuchs, B., Spring, S., Beisker, W., & Amann, R. (1997). Flow sorting of microorganisms for molecular analysis. *Applied and Environmental Microbiology*, *63*(11), 4223–31. Retrieved from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=168741&tool=pmcentrez&rendertype=abstract

Warnecke, F., & Hugenholtz, P. (2007). Building on basic metagenomics with complementary technologies. *Genome Biology*, *8*(12), 231. doi:10.1186/gb-2007-8-12-231

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., … Cheng, J.-F. (2011). Decontamination of MDA reagents for single cell whole genome amplification. *PloS One*, *6*(10), e26161. doi:10.1371/journal.pone.0026161

Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., … Cheng, J.-F. (2010). One bacterial cell, one complete genome. *PloS One*, *5*(4), e10314. doi:10.1371/journal.pone.0010314

Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., … Stepanauskas, R. (2009). Assembling the marine metagenome, one cell at a time. *PloS One*, *4*(4), e5299. doi:10.1371/journal.pone.0005299

Yilmaz, S., & Singh, A. K. (2012). Sinlge Cell Genome Sequencing. *National Institutes of Health*, *23*(3), 437–443. doi:10.1016/j.copbio.2011.11.018.SINGLE

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., … Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science (New York, N.Y.)*, *332*(6030), 714–7. doi:10.1126/science.1203163

Youssef, N. H., Blainey, P. C., Quake, S. R., & Elshahed, M. S. (2011). Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). *Applied and Environmental Microbiology*, *77*(21), 7804–14. doi:10.1128/AEM.06059-11

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W., & Church, G. M. (2006). Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, *24*(6), 680–6. doi:10.1038/nbt1214

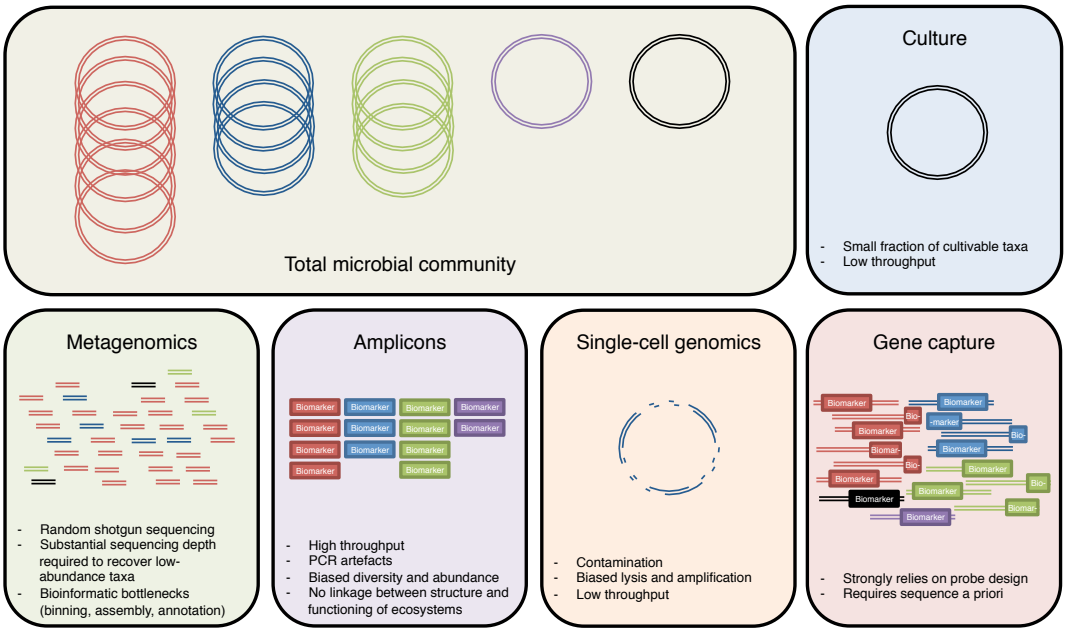
Zhang, L. I. N., Cui, X., Schmitt, K., Hubert, R., Navidit, W., & Arnheim, N. (1992). Whole genome amplification from a single cell: Implications for genetic analysis. *Proceedings of the National Academy of Sciences*, *89*(July), 5847–5851.

Table 1: Main candidate phyla (currently with no yet cultured representant). Based on the previously work of Rinke *et al.* (2013), McDonald *et al.* (2012), Riesenfield *et al.* (2004) and Hugenholtz (2002) and on the information available on Greengenes, Silva and RDP databases (in June 2014).
* References for sequenced genomes

Kingdom	Candidate phylum	First description in	Sequenced genomes	References
BACTERIA	**AD3**	Sandy surface soils	-	Zhou *et al.*, 2003
	BD1-5 group / GN02 [Gracilibacteria]	Guerrero Negro hypersaline microbial mat	5 unnamed draft genomes (metagenomics)	Ley *et al.*, 2006 Wrighton *et al.*, 2012*
	BH1	Near-boiling silica-depositing thermal springs	-	Blank *et al.*, 2002
	BRC1 / NKB19 [Hydrogenedentes]	Bulk soil and rice roots (BRC1 means Bacterial Rice Cluster)	-	Derakshani *et al.*, 2001
	CD12 / BHI80-139 [Aerophobetes]	-	-	Rinke *et al.*, 2013
	EM3 (former OP2)	Obsidian Pool, Yellowstone National Park	-	Hugenholtz *et al.*, 1998 Dunfield *et al.*, 2012*
	GN01	Guerrero Negro hypersaline microbial mat	-	Ley *et al.*, 2006
	GN04	Guerrero Negro hypersaline microbial mat	-	Ley *et al.*, 2006
	GOUTA4	Monochlorobenzene Contaminated Groundwater	-	Alfreider *et al.*, 2002
	KSB1	Sulfide-rich black mud from marine coastal environments	-	Tanner *et al.*, 2000s
	LD1	Anoxic Marine Sediments	-	Freitag *et al.*, 2003
	Marine Group A / SAR406 [Marinimicrobia]	Subsurface of Atlantic and Pacific oceans	-	Fuhrman *et al.*, 1993
	MVP-15	Suboxic freshwater pond	-	Briée *et al.*, 2007
	NC10	Aquatic microbial formations in flooded caves	*Methylomirabilis oxyfera* (metagenomics)	Holmes *et al.*, 2001 Ettwig *et al.*, 2010*
	OD1 / WWE3 [Parcubacteria]	Obsidian Pool, Yellowstone National Park (OD1 means OP11 Derived 1)	27 unnamed draft genomes (metagenomics) 1 unnamed circular genome (metagenomics)	Harris *et al.*, 2004 Wrighton *et al.*, 2012* Kantor *et al.*, 2013* Wrighton *et al.*, 2014*
	OP1 / KB1 group [Acetothermia]	Obsidian Pool, Yellowstone National Park	*Acetothermum autotrophicum* (metagenomics)	Hugenholtz *et al.*, 1998 Takami *et al.*, 2012*
	OP11 [Microgenomates]	Obsidian Pool, Yellowstone National Park	17 unnamed draft genomes (metagenomics) ZG1 draft genome (single-cell genomics)	Hugenholtz *et al.*, 1998 Wrighton *et al.*, 2012* Youssef *et al.*, 2011* Wrighton *et al.*, 2014*
	OP3 [Omnitrophica]	Obsidian Pool, Yellowstone National Park	-	Hugenholtz *et al.*, 1998
	OP8 [Aminicenantes]	Obsidian Pool, Yellowstone National Park	-	Hugenholtz *et al.*, 1998

Kingdom	Candidate phylum	First description in	Sequenced genomes	References
	OP9 / JS1 [Atribacteria]	Obsidian Pool, Yellowstone National Park	Draft genome 'Candidatus Caldatribacterium californiense' (single-cell genomics) Draft genome Caldatribacterium saccharofermentans' (metagenomics)	Hugenholtz *et al.*, 1998 Dodsworth *et al.*, 2013*
	Poribacteria	Marine sponge-associated	6 unnamed draft genomes (single-cell genomics)	Fieseler *et al.*, 2004 Siegl *et al.*, 2011* Kamke *et al.*, 2013*
	SBR1093	Activated sludge from an industrial wastewater treatment system	-	Layton *et al.*, 2000
	SC4	Arid soil from Arizona	-	Dunbar *et al.*, 2002
	SPAM	Alpine Soil in the Colorado Rocky Mountains (SPAM means SPring Alpine Meadow)	-	Lipson *et al.*, 2004
	SR1	Hydrocarbon-contaminated aquifer (SR means "Sulfur River")	1 unnamed draft genome (metagenomics) 1 unnamed circular genome (metagenomics) 1 unnamed draft genome (single-cell genomics)	Dojka *et al.*, 1998 Harris *et al.*, 2004 Campbell *et al.*, 2013* Kantor *et al.*, 2013* Wrighton *et al.*, 2014*
	TM6	Peat bog (TM mean Torf, Mittlere schicht)	1 unnamed draft genome (single-cell genomics)	Rheims *et al.*, 1996 Mclean *et al.*, 2013*
	TM7	Peat bog (TM mean Torf, Mittlere schicht)	5 drafts genomes (single cell genomics) 4 drafts genomes (metagenomics) 1 unnamed circular genome (metagenomics)	Rheims *et al.*, 1996 Marcy *et al.*, 2007* Podar *et al.*, 2007* Albertsen *et al.*, 2013* Kantor *et al.*, 2013*
	WPS-2	Wittenberg polluted soil	-	Nogales *et al.*, 2001
	WS1	Wurtsmith Air Force Base, Michigan	-	Dojka *et al.*, 1998
	WS2	Wurtsmith Air Force Base, Michigan	-	Dojka *et al.*, 1998
	WS3 [Latescibacteria]	Wurtsmith Air Force Base, Michigan	-	Dojka *et al.*, 1998
	WS6	Wurtsmith Air Force Base, Michigan	-	Dojka *et al.*, 1998
	WWE1 [Cloacimonetes]	Municipal Anaerobic Sludge Digester	*Cloacamonas acidaminovorans* (metagenomics)	Chouari *et al.*, 2005 Pelletier *et al.*, 2008*
	ZB3	Mesophilic Sulfide-Rich Spring	-	Elshahed *et al.*, 2003
ARCHAEA	**Korarchaeota**	Obsidian Pool, Yellowstone National Park	*Koarchaeum cryptophylum* (metagenomics)	Barns *et al.*, 1996 Elkins *et al.*, 2008*
	Nanoarchaeota	Submarine hot vent	-	Huber *et al.*, 2002s

Figure 1


Total microbial community
Culture
- Small fraction of cultivable taxa
- Low throughput
Metagenomics
- Random shotgun sequencing
- Substantial sequencing depth required to recover low-abundance taxa
- Bioinformatic bottlenecks (binning, assembly, annotation)
Amplicons
- High throughput
- PCR artefacts
- Biased diversity and abundance
- No linkage between structure and functioning of ecosystems
Single-cell genomics
- Contamination
- Biased lysis and amplification
- Low throughput
Gene capture
- Strongly relies on probe design
- Requires sequence a priori

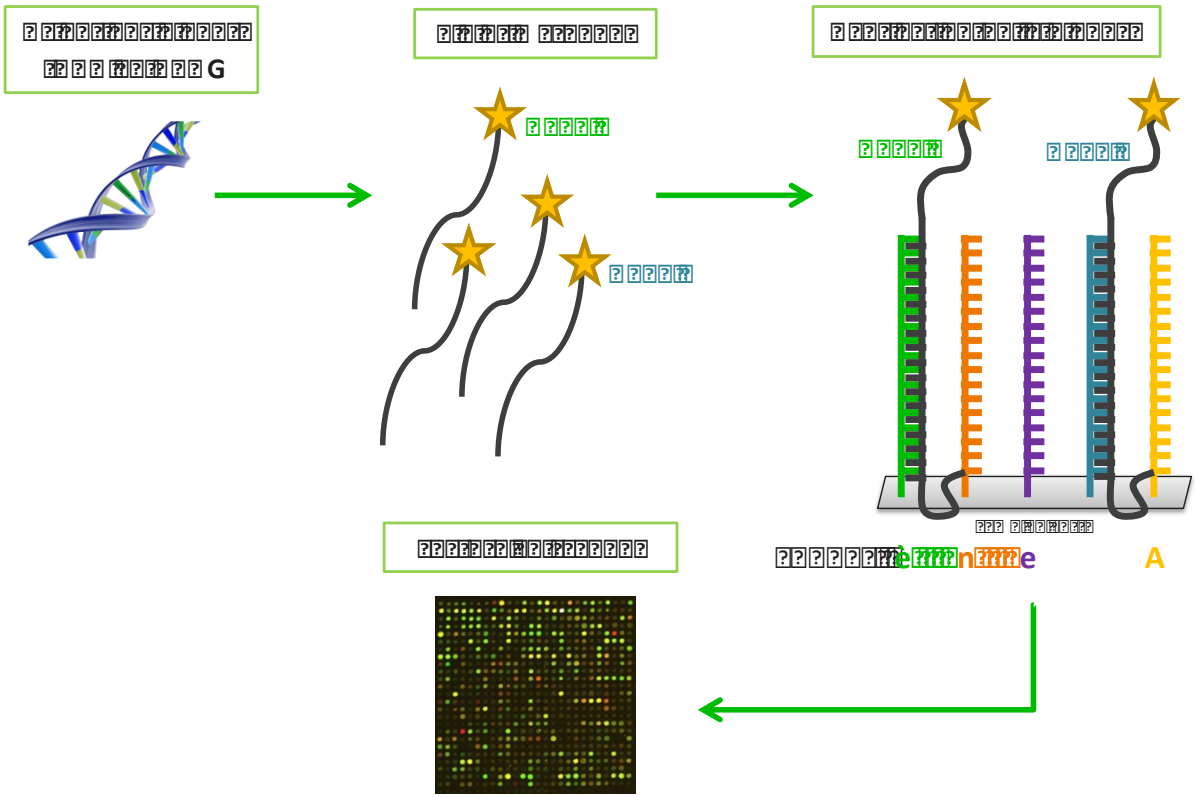


Figure 9. Principe des biopuces ADN.

Les cibles marquées à l'aide d'un fluorochrome s'hybrident spécifiquement avec les sondes qui leur sont complémentaires. L'analyse de l'image obtenue permet de déterminer quels sont les gènes présents dans l'échantillon.

Enfin, une dernière méthode de réduction de complexité peut être citée lorsque l'on s'intéresse à l'étude des populations microbiennes au sein des environnements, il s'agit des biopuces ADN (Zhou 2003 ; Gentry *et al.* 2006 ; Wagner *et al.* 2007).

2.2 Les biopuces ADN

Apparues au milieu des années 90 suite au séquençage des premiers génomes et issues de la rencontre des domaines tels que la microélectronique, les microsystèmes et la biologie, les biopuces ont été au départ mises au point pour l'étude simultanée de l'expression de tous les gènes d'un organisme (Schena *et al.* 1995). Les biopuces ADN ont connu un essor considérable ces 20 dernières années pour répondre aux problématiques de l'écologie microbienne.

2.2.1 Principe

Les biopuces ADN sont dérivées des techniques d'hybridation des acides nucléiques du *Southern blot* (Southern 1975) et du *Dot blot* (Kafatos *et al.* 1979), mais contrairement à ces dernières, l'hybridation est dite inverse puisque ce sont les sondes et non pas les cibles qui sont fixées sur un support solide (Ehrenreich 2006). Les sondes peuvent correspondre à de l'ADNg, des produits PCR, de l'ADNc ou encore à des oligodésoxyribonucléotides fixées sur une lame de verre, et elles vont agir comme des « hameçons » moléculaires en reconnaissant leurs cibles par complémentarité des bases. Les cibles sont des produits PCR, de l'ADNg, de l'ADNc ou encore des ARN. Plusieurs échantillons peuvent être hybridés simultanément en utilisant des marquages à l'aide de fluorophores différents (généralement les cyanines Cy3 et Cy5) (**Figure 9**). Une seule biopuce, sur laquelle une compartimentation physique est possible, peut donc permettre l'analyse de différents échantillons biologiques en une seule expérience ainsi que l'identification d'un grand nombre de séquences, puisque les formats actuels de biopuces permettent la fixation d'un million de sondes différentes. Suite à l'étape d'hybridation entre les sondes et les cibles fluorescentes, les duplex formés sont détectés à l'aide d'un scanner. Un faisceau laser va balayer toute la surface de la lame et exciter les fluorophores pour entraîner une émission de lumière. Les intensités lumineuses sont collectées puis transformées en signal électrique permettant d'évaluer quantitativement et qualitativement les échantillons hybridés.

Actuellement, les sondes oligodésoxyribonucléotidiques sont privilégiées du fait de leur plus faible coût et de la facilité de synthèse des sondes (Relógio *et al.* 2002). Il existe

deux principales technologies de fabrication des biopuces ADN : les biopuces dites *ex situ* nécessitant une préparation des sondes au préalable avant leur greffage sur le support solide, et les biopuces dites *in situ* pour lesquelles la synthèse des sondes est directement réalisée sur la lame de verre. Cette dernière technologie de synthèse de biopuce ADN est à privilégier pour les études haut-débit en écologie microbienne du fait de son coût moindre et de la possibilité de leur très haute densité de sondes (Dufva 2005 ; Kawasaki 2006 ; Dufva 2009).

2.2.2 Les biopuces ADN en écologie microbienne

La première biopuce ADN appliquée à l'écologie microbienne date de 1997. Elle était composée de neuf sondes ciblant le gène codant pour l'ARNr 16S et permettait l'identification de bactéries nitrifiantes (Guschin *et al.* 1997). Depuis, les biopuces ADN ont été utilisées dans de nombreuses études en écologie microbienne (Zhou & Thompson 2002 ; Zhou 2003 ; Wagner *et al.* 2007 ; Chan *et al.* 2013 ; Closek *et al.* 2014). Différentes catégories de biopuces ont ainsi été utilisées dont les « *Whole Genome Array* » (WGA) permettant de cibler dans son intégralité les gènes d'un ou plusieurs microorganismes et pouvant être utilisées notamment pour caractériser des souches ou des *consortia* isolés d'environnements complexes (Wu *et al.* 2004). Cependant, l'utilisation de ce type de biopuces pour l'étude *in situ* d'échantillons environnementaux est limitée en raison de l'importante complexité des communautés microbiennes composées en grande majorité de souches non caractérisées et pour lesquelles il n'existe aucune information de séquence. C'est pourquoi, l'utilisation de biopuces dites phylogénétiques (*Phylogenetic Oligonucleotide Array* ou POA) ou fonctionnelles (*Functional Gene Array* ou FGA) ciblant respectivement des biomarqueurs phylogénétiques et fonctionnels apparaît plus adaptée pour l'écologie microbienne (Gentry *et al.* 2006 ; Wagner *et al.* 2007).

2.2.2.a Biopuces phylogénétiques

Les différentes méthodes moléculaires développées et appliquées à l'écologie microbienne pour l'identification et la classification des communautés microbiennes appartenant au domaine des procaryotes (bactéries et archées) ciblent principalement la séquence codant pour l'ARNr 16S. Les régions les plus conservées de l'ADNr 16S permettent de déterminer des sondes assurant l'identification à des rangs taxonomiques supérieurs comme la famille, l'ordre ou la classe, alors que les régions plus variables peuvent discriminer les microorganismes, dans de nombreux cas, à des niveaux plus résolutifs comme le genre ou l'espèce (Huyghe *et al.* 2008). De plus, l'accessibilité à une multitude de séquences présentes

au sein des bases de données dédiées comme SILVA (Quast *et al.* 2013), Greengenes (McDonald *et al.* 2012) et RDP (Cole *et al.* 2013) permet d'affiner la détermination des sondes mais également de tester leur spécificité contre l'ensemble des séquences ADNr 16S disponibles ou caractéristiques de l'environnement étudié.

Dès lors de nombreuses biopuces phylogénétiques dédiées à l'écologie microbienne ont été mises au point avec des sondes présentant une taille comprise entre 18 et 25-mers. Elles ont été utilisées pour l'étude i) de groupes bactériens spécifiques à différents niveaux taxonomiques comme la division des *Acidobacteria* (Liles *et al.* 2010) ou le genre *Burkholderia* (Schönmann *et al.* 2009) ; ii) de communautés microbiennes d'environnements spécifiques comme des sols pollués par des solvants chlorés (Nemir *et al.* 2010), la rhizosphère du blé (Sanguin *et al.* 2009), le microbiote intestinal (Rajilić-Stojanović *et al.* 2009 ; 2012), ou encore de façon plus originale le métagénome de la cigarette (Sapkota *et al.* 2010) ; iii) de groupes fonctionnels particuliers comme les sulfato-réducteurs (Loy *et al.* 2002) ou les bactéries nitrifiantes (Kelly *et al.* 2005).

En parallèle, une biopuce phylogénétique plus généraliste nommée PhyloChip (Brodie *et al.* 2006), contenant environ 500 000 sondes de 25-mers et ciblant près de 9000 OTUs, a été développé et permet aujourd'hui de couvrir la quasi totalité de la diversité des communautés procaryotes répertoriée dans les bases de données. Cet outil a été utilisé pour étudier les communautés microbiennes issues d'environnements complexes : l'air intérieur des avions (Korves *et al.* 2013), des nappes phréatiques contaminées par du trichloroéthylène (Lee *et al.* 2012), des sols de prairies (DeAngelis *et al.* 2009 ; Cruz-Martínez *et al.* 2009 ; Delmont *et al.* 2011 ; He *et al.* 2012), les roches volcaniques (Kelly *et al.* 2010 ; 2011), des sédiments de rivières contaminés par des métaux lourds (Rastogi *et al.* 2011), les hautes couches de l'atmosphère (Smith *et al.* 2013), des coraux (Roder *et al.* 2014), ou encore des sols de l'Antarctique (Yergeau *et al.* 2009). Cependant, en raison d'une connaissance partielle des microorganismes de l'environnement, les informations obtenues par ces approches ne permettent d'identifier que les espèces pour lesquelles des séquences sont disponibles.

Les biopuces phylogénétiques ont démontré leur pertinence en écologie microbienne validée notamment par les nouvelles techniques de séquençage. En effet, des études ont montré une forte corrélation entre les résultats des biopuces et des NGS (Roh *et al.* 2010). Par exemple, l'utilisation de la HITChip (*Human Intestinal Tract Chip*) *versus* le pyroséquençage des régions V4 et V6 de l'ADNr 16S au sein d'échantillons de selles de patients âgés, a

montré une bonne corrélation des résultats au niveau du phylum (r = 0,94), de la classe (r = 0,93) ou encore de l'ordre (r = 0,94) (Claesson *et al.* 2009). Le même résultat a pu être retrouvé au niveau du microbiome humain, avec l'utilisation de la *Human Oral Microbe Identification Microarray* (HOMIM) (Preza *et al.* 2009) *versus* le pyroséquençage des régions V3-V5 de l'ADNr 16S, où une forte corrélation des résultats au niveau du phylum et du genre a pu être obtenue (Ahn *et al.* 2011). Cependant, même si les biopuces phylogénétiques apportent des informations précises sur la structure des communautés microbiennes au sein d'environnements complexes, elles ne permettent pas l'identification des capacités métaboliques. Ceci est d'autant plus problématique quand différents membres d'un même groupe de microorganismes présentent des capacités métaboliques différentes et qu'ils ne peuvent être différenciés sur la seule base de leurs signatures moléculaires d'ADNr 16S. Ainsi, l'étude des capacités métaboliques d'une communauté microbienne nécessite l'utilisation préférentielle de biopuces fonctionnelles (FGA) ciblant directement les gènes impliqués dans les processus métaboliques d'intérêt (He *et al.* 2011).

2.2.2.b Biopuces fonctionnelles

L'utilisation des biopuces fonctionnelles nécessite au préalable un choix des gènes cibles basé sur des critères précis. Ces gènes doivent i) coder pour une protéine clé dans la voie métabolique ciblée ; ii) présenter des régions suffisamment discriminantes pour assurer la détection spécifique du gène ciblé ; iii) être représentés par un maximum de séquences dans les bases de données permettant de couvrir un maximum de variants. Généralement, les sondes longues (50- à 70-mers) sont privilégiées pour la conception de telles biopuces fonctionnelles, puisqu'elles offrent une meilleure sensibilité tout en gardant une spécificité suffisante en relation avec la variabilité généralement rencontrée entre les gènes fonctionnels (Gentry *et al.* 2006). Il a été montré que l'utilisation de sondes de 50-mers permet de discriminer des séquences montrant moins de 88% d'identité tout en conservant une sensibilité suffisante ne nécessitant l'utilisation que de 5 à 10 ng d'ADNg seul ou 50 à 100 ng d'ADNg en mélange (Rhee *et al.* 2004). De tels seuils de détection correspondent à la mise en évidence de cibles ADN provenant seulement de 10 cellules ou de 0,03% à 5% des populations présentes dans une communauté bactérienne (Bodrossy *et al.* 2003 ; Peplies *et al.* 2004 ; Loy *et al.* 2005 ; Palmer *et al.* 2006 ; Gentry *et al.* 2006 ; Marcelino *et al.* 2006 ; Huyghe *et al.* 2008 ; Rajilić-Stojanović *et al.* 2009). En outre, il a été montré que pour des quantités comprises entre 1 à 100 ng d'ADNg (pur ou en mélange) une relation linéaire

pouvait être établie entre intensité du signal et quantité de cible hybridée, et permettre ainsi une analyse semi-quantitative (Wu *et al.* 2001).

La première biopuce fonctionnelle était composée d'environ 100 sondes construites à partir de produits PCR et ciblant différents gènes du cycle de l'azote (Wu *et al.* 2001). Depuis, l'utilisation de sondes oligonucléotidiques a permis, au cours de ces dix dernières années, la conception à haute densité et à moindre coût de biopuces fonctionnelles dédiées ciblant un ou plusieurs métabolismes particuliers, et d'autres plus généralistes s'intéressant aux nombreux gènes impliqués dans la plupart des réactions biochimiques et des cycles biogéochimiques. Les biopuces fonctionnelles spécifiques ont été élaborées pour répondre à des questions biologiques précises comme celles en relation avec la résistance à un antibiotique (Call *et al.* 2003), à la dégradation d'hydrocarbures aromatiques polycycliques (Terrat *et al.* 2010) et la résistance aux métaux (Rhee *et al.* 2004), la dégradation de solvants chlorés (Dugat-Bony *et al.* 2012a), à la dégradation des polychlorobyphényls (PCBs) (Denef *et al.* 2003), du benzène (Iwai *et al.* 2007 ; 2008), à des facteurs de virulence bactérienne (Jaing *et al.* 2008 ; Miller *et al.* 2008 ; Lee *et al.* 2013), au microbiote intestinal (Tu *et al.* 2014b), au cycle du méthane (Bodrossy *et al.* 2003 ; Stralis-Pavese *et al.* 2011), de l'azote (Taroncher-Oldenburg *et al.* 2003 ; Tiquia *et al.* 2006 ; Ward *et al.* 2007 ; Duc *et al.* 2009 ; Ward & Bouskill 2011) et du soufre (Rinta-Kanto *et al.* 2011). Plus récemment une biopuce fonctionnelle dédiée à l'étude des procédés de bioremédiation, la « BiodegPhyloChip » a été mise au point pour détecter 1057 gènes impliqués dans la dégradation de 133 polluants et mettre ainsi en avant les capacités métaboliques des communautés microbiennes de différents sites contaminés (Pathak *et al.* 2011). Cependant, même si cet outil permet de couvrir une large gamme d'informations sur la dégradation de nombreux polluants, les sondes déterminées ne ciblent pas tous les gènes impliqués dans le processus de biodégradation ni même l'ensemble des variants de chaque gène.

Des biopuces fonctionnelles plus généralistes ont été mises au point et sont actuellement disponibles comme la GeoChip (He *et al.* 2011) qui a subi différentes évolutions, permettant de passer d'une première biopuce ciblant 2402 gènes (GeoChip 1.0) (He *et al.* 2007) à 141 995 (GeoChip 4.0) (Tu *et al.* 2014c). Les différentes versions de la GeoChip sont actuellement les biopuces fonctionnelles les plus utilisées en écologie microbienne pour caractériser la diversité fonctionnelle des environnements complexes. Leurs applications ont été diverses avec, par exemple, l'étude de sites pollués par des fuites de pétrole (Beazley *et al.* 2012) ou des hydrocarbures aromatiques polycycliques (Ding *et al.*

2012) ; des sols de mangroves (Bai *et al.* 2013) ou de prairies (Yang *et al.* 2014) ; ou encore des boues activées de stations d'épuration (Wang *et al.* 2014a).

Cependant, même si ces outils ciblent une large gamme de gènes impliqués dans tous les processus métaboliques globaux, leurs applications restent limitées à l'image des biopuces phylogénétiques, par leur incapacité à appréhender la diversité fonctionnelle encore non décrite.

3. Stratégies et outils pour la sélection de sondes oligonucléotidiques

A l'exception des techniques de cellule isolée, les approches précédemment citées reposent sur l'utilisation d'amorces (amplicons) ou de sondes (biopuces et capture de gènes) oligonucléotidiques. C'est pourquoi l'efficacité de ces méthodes est entièrement dépendante de l'utilisation d'oligonucléotides de haute qualité.

Ainsi, face à l'augmentation constante du nombre de séquences déposées dans les bases de données, les stratégies de détermination de sondes oligonucléotidiques doivent être de plus en plus performantes et les algorithmes entièrement repensés.

3.1 Détection de séquences inconnues : stratégies de *design* de sondes exploratoires

Grâce à l'essor spectaculaire des techniques moléculaires, les chercheurs ont littéralement changé leur manière d'appréhender le monde microbien. De nombreux outils qualifiés de « haut-débit » sont apparus et avec eux, la possibilité de produire et de traiter des quantités de données gigantesques, jusqu'alors inaccessibles, et pourtant indispensables pour comprendre le fonctionnement des écosystèmes. Parmi les technologies les plus utilisées actuellement, les biopuces ADN oligonucléotidiques sont considérées comme des outils de choix pour répondre aux problématiques d'écologie microbienne (Zhou 2003 ; Gentry *et al.* 2006 ; Wagner *et al.* 2007). En effet, grâce à elles, il est possible de caractériser dans un échantillon environnemental, la présence et/ou l'expression de plusieurs milliers de gènes et cela au cours d'une même expérience.

La pertinence d'une expérimentation biopuce ADN pour l'étude des communautés microbiennes est entièrement dépendante des sondes sélectionnées. Outre les caractéristiques obligatoires qu'elles doivent présenter (sensibilité, spécificité et uniformité), leur utilisation pour des applications en écologie microbienne nécessite un caractère supplémentaire, dit « exploratoire », pour permettre la détection de l'ensemble des populations connues mais également, celles encore jamais décrites. Or, actuellement, la plupart des stratégies de détermination de sondes proposées reposent uniquement sur les informations de séquences présentes dans les bases de données internationales (Lemoine *et al.* 2009). Les sondes sélectionnées ne ciblent, par conséquent, que les microorganismes pour lesquels des séquences sont disponibles.

Afin d'avoir une vue globale des stratégies bioinformatiques disponibles et applicables aux problématiques d'écologie microbienne pour la détermination de sondes, nous avons réalisé un état de l'art à partir de données bibliographiques. Cette partie est présentée sous forme d'une revue scientifique, publiée dans le journal « *Environmental Microbiology* », et offre une présentation détaillée des nouvelles approches de détermination de sondes exploratoires.

Article n°2

Detecting unknown sequences with DNA microarrays: explorative probe design strategies.

environmental microbiology

Environmental Microbiology (2012) **14**(2), 356–371 doi:10.1111/j.1462-2920.2011.02559.x

Minireview

Detecting unknown sequences with DNA microarrays: explorative probe design strategies

Eric Dugat-Bony,[1,2] Eric Peyretaillade,[1,2] Nicolas Parisot,[1,2] Corinne Biderre-Petit,[1,2] Faouzi Jaziri,[3,4] David Hill,[3,4] Sébastien Rimour[1,2] and Pierre Peyret[1,2]*

[1]*Clermont Université, Université Blaise Pascal, Laboratoire Microorganismes: Génome et Environnement, BP 10448, F-63000, Clermont-Ferrand.*
[2]*UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France.*
[3]*Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand.*
[4]*UMR CNRS 6158, LIMOS, F-63173 Aubière.*

Summary

Designing environmental DNA microarrays that can be used to survey the extreme diversity of microorganisms existing in nature, represents a stimulating challenge in the field of molecular ecology. Indeed, recent efforts in metagenomics have produced a substantial amount of sequence information from various ecosystems, and will continue to accumulate large amounts of sequence data given the qualitative and quantitative improvements in the next-generation sequencing methods. It is now possible to take advantage of these data to develop comprehensive microarrays by using explorative probe design strategies. Such strategies anticipate genetic variations and thus are able to detect known and unknown sequences in environmental samples. In this review, we provide a detailed overview of the probe design strategies currently available to construct both phylogenetic and functional DNA microarrays, with emphasis on those permitting the selection of such explorative probes. Furthermore, exploration of complex environments requires particular attention on probe sensitivity and specificity criteria. Finally, these innovative probe design approaches require exploiting newly available high-density microarray formats.

Received 31 March, 2011; accepted 4 July, 2011. *For correspondence. E-mail pierre.peyret@univ-bpclermont.fr; Tel. (+33) 473405139; Fax (+33) 473407670.

Introduction

The microbial world represents the most important and diverse group of organisms living on earth (Whitman *et al*., 1998; Curtis *et al*., 2002), comprising most of the diversity of the three domains of life defined by Woese and colleagues (1990): *Archaea*, *Bacteria* and *Eucarya*. Furthermore, these organisms are widely distributed across many environmental habitats, even the most extreme. Their numerous enzymatic machineries have allowed them to adapt to almost every ecological niche and take advantage of any environmental condition (Øvreås, 2000; Guerrero and Berlanga, 2006). Despite our increasing knowledge of the role of microorganisms in ecosystem functioning, our current vision of the microbial world is still incomplete and several issues remain unsolved. This is partially explained (i) by the tremendous diversity of the genes and metabolisms of the existing species but also of ecological niches and (ii) by technological limits such as our inability to culture the majority of microorganisms (Amann *et al*., 1995; Pace, 1997).

Because of this huge microbial biocomplexity, high-throughput molecular tools allowing simultaneous analyses of existing populations are greatly needed (Torsvik and Øvreås, 2002; Xu, 2006). Massive sequencing based on next-generation sequencing (NGS) technologies and microarrays are currently the most promising and complementary approaches to address these tasks (Claesson *et al*., 2009; Roh *et al*., 2010; van den Bogert *et al*., 2011). Using NGS, two specific strategies can be applied: metagenomics, which refers to the study of the collective genomes in a given environmental community and the 16S rDNA amplicon sequencing approach. In principle, these methods enable: (i) access to the wide diversity of microbial communities, (ii) identification of unknown microorganisms and (iii) the potential to link structure to functions (Simon and Daniel, 2009). Some limitations of metagenomics, however, have been demonstrated: for example, the huge difficulty of managing large amounts of sequence data, or the short sequence read length (400–500 bases maximum with 454 FLX Titanium instrument from Roche), which complicates contigs assembling, or the sequencing errors caused by NGS

technologies (Roh *et al*., 2010). Furthermore, Quince and colleagues (2008) estimated that detecting 90% of the richness in some hyperdiverse environments could require tens of thousands of times the current sequencing effort, which is inconceivable. Oligonucleotide microarray technologies have, however, been widely used for gene detection and gene expression quantification, and more recently, were adapted to profiling environmental communities in a flexible and easy-to-use manner (Zhou, 2003; Wagner *et al*., 2007). These approaches can monitor the presence, or the expression, of thousands of genes, combining qualitative and quantitative aspects in only one experiment (Tiquia *et al*., 2004; Marcelino *et al*., 2006; Dugat-Bony *et al*., 2011). Furthermore, this technology appears well adapted to multi-sample comparison. Although several whole-genome arrays have been developed in the last few years, phylogenetic oligonucleotide arrays (POAs), targeting the 16S rRNA genes, as well as functional gene arrays (FGAs), targeting key genes encoding enzymes involved in metabolic processes, are the two major approaches to assess diversity of microbial communities in the environment (Wagner *et al*., 2007). Currently, the most comprehensive tools developed are the high-density PhyloChip, with nearly 500 000 oligonucleotide probes to almost 9000 operational taxonomic units (Brodie *et al*., 2006), and the GeoChip 3.0 with ~ 28 000 probes covering approximately 57 000 gene variants from 292 functional gene families (He *et al*., 2010). Whereas microarrays were demonstrated as being sufficiently sensitive, with detection of sequences representing genomic material from 0.05% to 5% of the total environmental community (Bodrossy *et al*., 2003; Peplies *et al*., 2004; Loy *et al*., 2005; Gentry *et al*., 2006; Marcelino *et al*., 2006; Palmer *et al*., 2006; Huyghe *et al*., 2008), these methods require a sequence *a priori* to determine probes and hence allow surveys only of microorganisms with available sequences in public databases (Chandler and Jarrell, 2005; Wagner *et al*., 2007).

The main problem that must be faced to construct oligonucleotide microarrays dedicated to microbial ecology is the probe design step. Indeed, environmental microarrays often require this step to be manually performed. Although numerous general probe design programmes are currently freely accessible for academics [for recent reviews see Lemoine and colleagues (2009)], only few may be useful for microbial ecology applications and are listed in Table 1. This review aims to show how probe design strategies can avoid the limitation of sequence availability and make possible the detection of previously uncharacterized microbial populations present in nature. We emphasize various recent methods combining the use of both degenerate and non-degenerate oligonucleotide probes to target either 16S rRNA markers, or new proteic variants. In conclusion, we highlight other procedures and

Table 1. Appropriate probe design software for microbial ecology studies.

Software	Applications in microbial ecology	Accessibility and user interface	URL	Reference
ARB	POA	Downloadable, standalone GUI (L, M)	http://www.arb-home.de/	Ludwig *et al*. (2004)
PRIMROSE	POA	Downloadable, GUI (L, W, M)	http://www.bioinformatics-toolkit.org/Primrose/index.html	Ashelford *et al*. (2002)
ORMA	POA, FGA	Matlab Script	Upon request	Severgnini *et al*. 2009)
PhylArray	POA	Web Interface	http://g2im.u-clermont1.fr/serimour/phylarray/	Militon *et al*. (2007)
HPD	FGA	Downloadable, standalone GUI (W)	Not available	Chung *et al*. (2005)
ProDesign	FGA	Web Interface	http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html	Feng and Tillier (2007)
HiSpOD	FGA, WGA	Web Interface	http://fc.isima.fr/~g2im/hispod/	Dugat-Bony *et al*. (2011)
Metabolic Design	FGA	Downloadable from a website, GUI (W)	ftp://195.221.123.90/	Terrat *et al*. (2010)
CommOligo (v 2.0)	FGA, WGA	Downloadable, standalone GUI (W)	http://ieg.ou.edu/software.htm	Li *et al*. (2005)
OligoWiz (v 2.0)	FGA, WGA	Downloadable client programme, GUI (L, W, M)	http://www.cbs.dtu.dk/services/OligoWiz2/	Wernersson and Nielsen (2005)
ROSO	FGA, WGA	Web interface or standalone GUI (S, W, M) upon request	http://pbil.univ-lyon1.fr/roso/Home.php	Reymond *et al*. (2004)
ArrayOligoSelector	FGA, WGA	Downloadable, command line (L)	http://arrayoligosel.sourceforge.net/	Bozdech *et al*. (2003)
OligoArray (v 2.1)	FGA, WGA	Downloadable, command line (L)	http://berry.engin.umich.edu/oligoarray2_1/	Rouillard *et al*. (2003)
OligoPicker	FGA, WGA	Downloadable, command line (L)	http://pga.mgh.harvard.edu/oligopicker/	Wang and Seed (2003)
PROBEmer	POA, FGA, WGA	Web Interface	Not available	Emrich *et al*. (2003)
YODA	FGA, WGA	Downloadable, standalone GUI (L, W, M)	Not available	Nordberg (2005)
ProbeSelect	WGA	Available upon request, command line (L)	Not available	Li and Stormo (2001)

FGA, functional gene array; GUI, graphical user interface; L, Linux; M, MacOS; POA, phylogenetic oligonucleotide array; S, SunOS; W, Windows; WGA, whole-genome array.

limitations that must be circumvented, to improve microarray development in terms of specificity and sensitivity.

General criteria for probe design

In silico probe design is one of the most critical step for microarray experiments because the selected oligonucleotide probe set will have to combine: (i) sensitivity (e.g. probes should detect low abundance targets in complex mixtures), (ii) specificity (e.g. probes should not cross-hybridize with non-target sequences) and (iii) uniformity (e.g. probes should display similar hybridization behaviour) (Loy and Bodrossy, 2006; Wagner *et al*., 2007). According to Lemoine and colleagues (2009), this process requires dealing with many parameters and currently available probe design programmes differ in the choice of criteria that are considered to select the best probe set (Table 2).

Sensitivity

The sensitivity generally increases with probe length, as the binding energy for longer probe-target hybrid complexes is typically higher and hybridization kinetics are irreversible (Hughes *et al*., 2001; Relogio *et al*., 2002; Letowski *et al*., 2004). For example, probes of 60 mers can detect targets with eightfold higher sensitivity than those of 25 mers (Chou *et al*., 2004). However, their threshold for differentiation is at 75–90% sequence similarity (Kane *et al*., 2000; Taroncher-Oldenburg *et al*., 2003; Tiquia *et al*., 2004), which indicates a poor specificity (Li *et al*., 2005). In contrast, short oligonucleotide probes are more specific, allowing discrimination of single nucleotide polymorphisms under optimal conditions, but at the cost of reduced sensitivity (Relogio *et al*., 2002). Furthermore, the formation of stable secondary self-structures like stem-loops, hairpins and probe-to-probe dimerization by the probes or their targets is another crucial factor that must be considered to minimize loss of microarray sensitivity. However, despite a good knowledge of the thermodynamic properties of nucleic acid duplex formation and dissociation in solution (SantaLucia *et al*., 1996) and the availability of several algorithms like Mfold (Zuker, 2003) or Hyther (Bommarito *et al*., 2000) for their accurate prediction, these calculations should be treated cautiously in the microarray context due to the limited knowledge on the thermodynamics of hybridization at solid–liquid interfaces (Pozhitkov *et al*., 2006; 2007).

Specificity

The specificity of microarray hybridization is one of the main effectors of the result quality (Kane *et al*., 2000; Evertsz *et al*., 2001; Koltai and Weingarten-Baror, 2008). Therefore, it is crucial that oligonucleotide probes must be unique with respect to all non-target sequences. To check probe specificity, software usually use results produced by algorithms such as BLAST or suffix array method, to search for cross-hybridization against databases constructed in accordance with the microarray application. In this step, potential cross-hybridization prediction are usually based on Kane's recommendations (probe should not have a total percent identity > 75–80% with a non-target sequence, or contiguous stretches of identity > 15 nt with a non-target sequence) (Kane *et al*., 2000) or thermodynamics calculations (duplex's stability between the probe and the non-target sequence). Moreover, low-complexity regions such as those containing long homopolymers may also contribute to affect probe specificity and must therefore be avoided for probe design (Wang and Seed, 2003; Leparc *et al*., 2009).

Uniformity

Because microarray technology relies on the simultaneous hybridization of many probes under the same conditions (salt concentration, temperature, etc.), it is important to ensure that the selected probes have thermodynamic behaviours as uniform as possible (Loy and Bodrossy, 2006; Wagner *et al*., 2007). The easiest way to achieve this is to select probes with homogeneous structural properties such as probe length, G + C content, melting temperature (T_m) or binding capacities (ΔG).

Characterization of environmental species with POAs

The classical way to characterize members of complex bacterial communities relies on the small subunit ribosomal RNA gene (16S rRNA) analysis. This target is particularly well adapted to phylogenetic studies as it contains highly conserved and variable moieties permitting reliable and detailed bacterial classification. Moreover, the advent of many PCR-based approaches, as well as sequencing projects, has led to the explosion of 16S rRNA gene sequences now available in major specialized sequence repositories, such as SILVA (Pruesse *et al*., 2007), Greengenes (DeSantis *et al*., 2006) and the Ribosomal Database Project (RDP) (Cole *et al*., 2009).

In order to rapidly survey prokaryotic communities present in complex environments high-throughput tools have been developed, such as POAs using the SSU rRNA biomarker (Wilson *et al*., 2002; Brodie *et al*., 2006; Palmer *et al*., 2006; DeSantis *et al*., 2007). The main obstacle in designing a POA, however, is potential cross-hybridization. In many cases, the 16S rRNA genes of the type species are too conserved to allow the design of

Table 2. Comparison of probe design software features.

Software	Probe length (nt)	Secondary structure	Low-complexity	GC content	T_m	ΔG	Degenerate probes	Cross-hybridization assessment	Database for specificity test
ARB	Fixed by the user (10–100)	No	No	Yes	Yes	No	No	Local alignment and thermodynamic calculations	ARB-Silva Database
PRIMROSE	Fixed by the user (3–100)	No	No	No	No	No	Yes	ND	RDP-II Database
ORMA	Fixed by the user	No	Yes	No	Yes	No	Yes	No	No
PhylArray	Fixed by the user (20–70)	No	No	Yes	Yes	No	Yes	BLAST and Kane's specifications	Custom non-redundant SSU rRNA database (95 Mo)
HPD	Fixed by the user (20–70)	Yes	No	Yes	Yes	Yes	No	BLAST and Kane's specifications	Input sequence dataset
ProDesign	Fixed by the user (20–70)	Yes	Yes	Yes	Yes	Yes	No	Spaced seed hashing and Kane's specifications	Input sequence dataset
HiSpOD	Fixed by the user	No	Yes	Yes	Yes	No	Yes	BLAST and Kane's specifications	EnvExBase (10Go) Complete CDS Database
Metabolic Design	Fixed by the user	No	No	No	No	No	Yes	BLAST and Kane's specifications	EnvExBase (10Go) Complete CDS Database
CommOligo (v 2.0)	Fixed by the user	Yes	Yes	Yes	Yes	No	No	Global alignment, thermodynamic calculations and Kane's specifications	Input sequence dataset
OligoWiz (v 2.0)	Fixed by the user	Yes	No	No	Yes	No	No	BLAST, Kane's specifications and thermodynamic calculations	Single organism genome
ROSO	Fixed by the user	Yes	Yes	Yes	Yes	Yes	No	BLAST	External fasta file (typically Single organism genome)
ArrayOligoSelector	Fixed by the user	Yes	Yes	Yes	No	No	No	BLAST and thermodynamic calculations	External fasta file (typically single organism genome)
OligoArray (v 2.1)	Fixed by the user (15–75)	Yes	Yes	Yes	Yes	No	No	BLAST and thermodynamic calculations	External fasta file (typically single organism genome)
OligoPicker	Fixed by the user (20–100)	Yes	Yes	No	Yes	No	No	BLAST	Input sequence dataset or external fasta file (typically single organism genome)
PROBEmer	Fixed by the user	Yes	No	Yes	Yes	Yes	No	Suffix array approach	RDP (v 8.1), single organism genome or external fasta file
YODA	Fixed by the user	Yes	Yes	Yes	Yes	No	No	BLAST and Kane's specifications	External fasta file (typically single organism genome)
ProbeSelect	Fixed by the user	Yes	Yes	No	No	Yes	No	Suffix array approach and thermodynamic calculation	Single organism genome

ND, not determined.

discriminatory probes (Bae and Park, 2006). To circumvent this problem, a hierarchical design allows probing for microbial taxa at different phylogenetic levels (Huyghe *et al.*, 2008; Liles *et al.*, 2010), providing information on the presence or absence of the branches and the twigs on the Tree of Life.

Probe design for POA

Both fully automated software and manual approaches have been developed to design POAs, taking into account the main criteria for efficient probe design, which are sensitivity and specificity. Currently, three programmes have been developed to work with structured data for retrieving and analysing sequences from dedicated databases and to operate a phylogenetic probe design targeting the 16S rRNA gene.

The first programme is the Probe Design tool included in the ARB programme package (Ludwig *et al.*, 2004) which is commonly used to select 10–100 mer oligonucleotides. The first step in the programme consists of the target group selection. Second, the algorithm identifies unique sequence stretches that could serve as target sites, and subsequently returns a sorted list of potential oligonucleotides. Third, the suggested probes can be matched against all sequences in the database using the Probe Match software programme. ARB also proposes different sets of predefined probes, each targeting distinct phylogenetic groups. It has been widely used to develop low-density custom-made POAs, containing up to a few hundred oligonucleotide probes. These probes usually target either restricted microorganism groups known to perform a specific metabolism (Loy *et al.*, 2002; Kelly *et al.*, 2005; Franke-Whittle *et al.*, 2009), or belonging to a specific taxon (Castiglioni *et al.*, 2004; Lehner *et al.*, 2005; Loy *et al.*, 2005; Kyselkova *et al.*, 2008; Schonmann *et al.*, 2009; Liles *et al.*, 2010), or living in a habitat/ecosystem of particular interest (Neufeld *et al.*, 2006; Sanguin *et al.*, 2009). To illustrate this purpose, Sanguin and colleagues (2009) identified multiple changes in rhizobacterial community composition associated with the decline of take-all disease of wheat caused by the soil-borne fungus *Gaeumannomyces graminis* by using a taxonomic 16S rRNA-based microarray targeting both *Bacteria*, *Archaea* and the OP11 and OP2 candidate divisions. ARB has also been used to construct phylogenetic microarrays based on other biomarkers such as protein coding genes (Bodrossy *et al.*, 2003; Duc *et al.*, 2009).

The second programme is the PRIMROSE programme (Ashelford *et al.*, 2002), which uses standard or custom databases, and allows the design of degenerate probes. Initially, a multiple alignment is produced using all the different sequences representing a given taxon. Every probe is subsequently tested against all the sequences of the initial database, to characterize potential cross-hybridizations and to verify good coverage of the targeted taxon. Although this tool was developed to identify both phylogenetic probes and primers, it has been mainly applied to PCR-based and FISH (fluorescent *in situ* hybridization) approaches (Rusch and Amend, 2004; Yu *et al.*, 2005; Feldhaar *et al.*, 2007; Boeckaert *et al.*, 2008; Klitgaard *et al.*, 2008; Muhling *et al.*, 2008; Gittel *et al.*, 2009; Fraune *et al.*, 2010; Bers *et al.*, 2011). Few applications of POAs using PRIMROSE have been reported. Blaskovic and Barak (2005) reported the development of a user-friendly chip to specifically detect tick-borne bacteria responsible of human and animal diseases.

Nevertheless, neither of these two applications is built specifically for the determination of discriminating positions within a set of very similar sequences. The third programme is ORMA (Oligonucleotide Retrieving for Molecular Applications), which represents a good alternative solution (Severgnini *et al.*, 2009). This programme designs and selects oligonucleotide probes for molecular application experiments on sets of highly similar sequences. Although it was first applied to the design of probes targeting 16S rRNA genes, this software can be used on any set of highly correlated sequences, such as new potential phylogenetic biomarkers. Using this programme, Candela and colleagues (2010) designed the HTF-Microbi.Array allowing high taxonomic level fingerprinting of the human intestinal microbial community.

In parallel, other computational approaches not implemented under fully automated software were developed to design high-density POAs (thousands of oligonucleotide probes) allowing a comprehensive screening for all known bacterial or archaeal taxa with a single microarray (Wilson *et al.*, 2002; DeSantis *et al.*, 2007). These approaches rely on sophisticated algorithms for the design of a multitude of probes and for the analyses of highly complex hybridization patterns. The best example is the PhyloChip developed by Brodie and colleagues (2006), which contains 500 000 probes based on the Affymetrix GeneChip platform. This tool is able to simultaneously identify thousands of taxa present in an environmental sample and has been applied to characterize prokaryotic communities from ecosystems such as urban atmosphere (Brodie *et al.*, 2007), grassland soils (Cruz-Martinez *et al.*, 2009; DeAngelis *et al.*, 2009), Antarctic soils (Yergeau *et al.*, 2009), mining-impacted soils (Rastogi *et al.*, 2010a,b), metal-contaminated river sediments (Rastogi *et al.*, 2011), terrestrial volcanic glasses (Kelly *et al.*, 2010), rhizosphere of potato (Weinert *et al.*, 2011), citrus leaf (Sagaram *et al.*, 2009), endotracheal aspirates from patients colonized by *Pseudomonas aeruginosa* (Flanagan *et al.*, 2007), and pearly eyed thrasher eggs (Shawkey *et al.*, 2009). Recently, due to increased interest in microbes of human and animal

gastrointestinal tracts, a number of high-density microarrays were also developed to study the composition and activity of intestinal microbiota (Palmer *et al*., 2006; Paliy *et al*., 2009; Rajilic-Stojanovic *et al*., 2009).

The main limitation of all the strategies proposed for 16S rRNA probe design is that they only ensure the survey of known microorganisms with available sequences in public databases. Unfortunately, the vast majority of microbial species is still unidentified and, therefore, is not represented by sequences in public ribosomal rRNA databases. A major challenge for the future is improvement of microarray technology to, in part, rely on new strategies for the design of explorative probes targeting sequences, which have not yet been described.

Explorative probe design strategies for POA

The 'multiple probe concept' consists of using several probes targeting an organism at similar and different phylogenetic/taxonomic levels. Designing probes using this concept significantly reduces the risk of misidentification, and often allows discrimination of bacteria down to the species level (Ludwig *et al*., 1998; Loy and Bodrossy, 2006; Schliep and Rahmann, 2006; Huyghe *et al*., 2008; Schonmann *et al*., 2009; Liles *et al*., 2010). Arrays constructed using this concept may ensure the detection of unknown taxa using probes defined from known higher phylogenetic levels. Because such probes are strictly complementary to available sequences, however, they do not harbour the explorative power to detect microorganisms with uncharacterized phylogenetic signatures.

Currently, the only software dedicated to POAs offering the possibility of designing explorative probes, is the PhylArray programme (Militon *et al*., 2007) This algorithm generates 16S rRNA probes to globally monitor known and unknown bacterial communities in complex environments. The first step in the design is the extraction of all available sequences corresponding to a given taxon from a custom 16S rRNA curated database. Second, a multiple sequence alignment is performed using the ClustalW algorithm (Thompson *et al*., 1994). Third, a degenerate consensus sequence is produced taking into account sequence variability at each position, which allows the selection of degenerate probes. Finally, all combinations from each degenerate probe are checked for cross-hybridization against the 16S rRNA database. Among the combinations derived from each degenerate probe, some correspond to sequences not previously included in public databases (Fig. 1). They should, therefore, allow for the exploration of the as yet undescribed fraction of environmental microbial communities. Moreover, comparative experimental evaluations indicate that probes designed with PhylArray yield a higher sensitivity and specificity than those designed with the PRIMROSE and ARB strategies (Militon *et al*., 2007). Recently, a microarray designed with the PhylArray strategy has been employed to evaluate the bacterial diversity in two different soils (Delmont *et al*., 2011). The authors highlighted the significant influence of several parameters like sampling depth or DNA extraction protocols on the biodiversity estimation.

Detection of functional signatures for FGA design

Assessing the metabolic potential of microorganisms in natural ecosystems is an interesting goal in microbial ecology. In fact, some authors estimate that individual environmental samples, like soil, may contain between 10^3 and 10^7 different bacterial genomes (Curtis *et al*., 2002; Gans *et al*., 2005), each of them harbouring thousands of genes. In this context, high-density oligonucleotide FGAs provide the best high-throughput tools to access this tremendous genetic content (He *et al*., 2008). GeoChips, composed of 50 mer probes designed with CommOligo (Li *et al*., 2005), are currently the most comprehensive FGAs. Indeed, these microarrays have evolved over several generations and now target key genes involved in most microbial functional processes such as carbon, nitrogen, phosphorus and sulfur cycles, energy metabolism, antibiotic resistance, metal resistance and organic contaminant degradation (Rhee *et al*., 2004; He *et al*., 2007; 2010; 2011). However, being able to encompass the full diversity of gene family sequences encountered in nature, described in databanks or not, is still one of the most difficult challenges for the future. Most FGAs described to date only monitor sequences available in databases and, therefore, cannot appraise the unknown part of the microbial gene diversity present in complex environments. A more extensive coverage of the probe set is, therefore, crucial and designing explorative probes represents a pertinent and essential approach.

Characterization of new functional signatures from nucleic sequence alignment

Many probe design programmes are currently freely accessible for academics [for recent reviews see Lemoine and colleagues (2009)]. Most of them were developed for use on single-genome datasets, and hence, are limited to the determination of probes targeting specific gene sequences (Table 2). In contrast, few strategies offer the opportunity to design probes allowing a broad coverage of multiple sequence variants for a given gene family.

With the availability of more and more sequences corresponding to functional genes (complete genome sequencing and environmental studies from specific functional markers), new programmes have been developed in the last decade taking into account this wide diversity. Hierarchical Probe Design (HPD) was the first programme

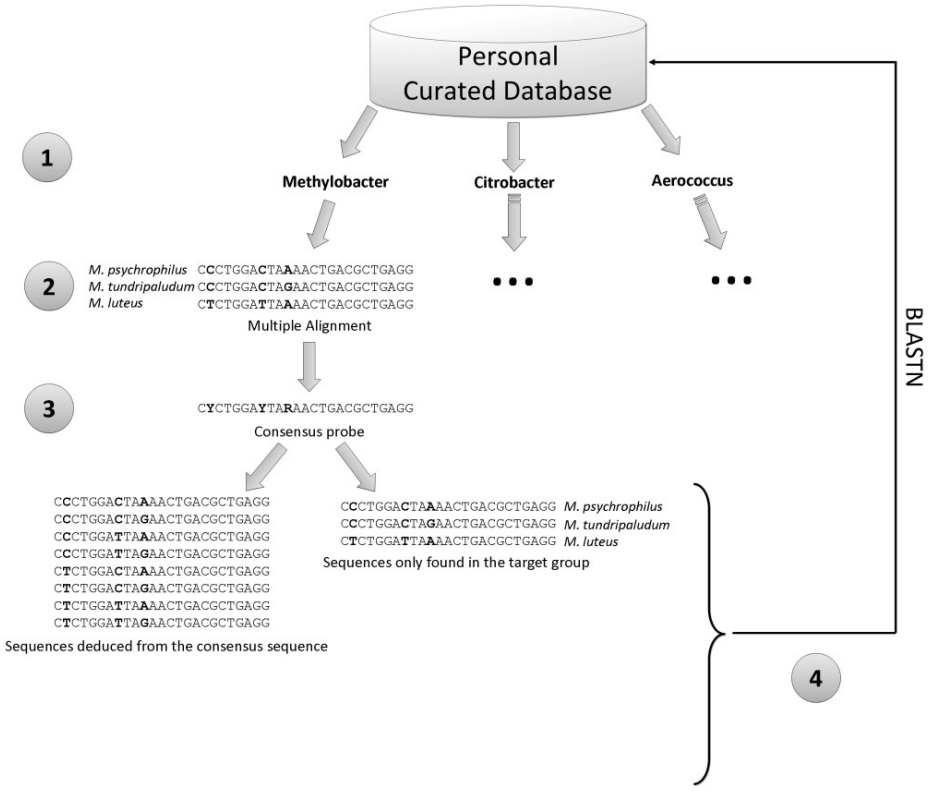


Fig. 1. PhylArray programme workflow. PhylArray programme is composed of four steps: (i) sequence extraction for each taxon, (ii) multiple sequence alignment, (iii) degenerate consensus sequence production and probe selection and (iv) specificity tests against the 16S rRNA database.

dedicated to functional oligonucleotide determination based on the concept of cluster-specific probes (Chung *et al.*, 2005). The first step of the programme consists of the alignment and hierarchical clustering of input sequences in order to generate all possible candidate probes. The optimal probe set is subsequently determined according to probe quality criteria, including cluster coverage, specificity, GC content and hairpin energy. Although this tool is not explorative, it automatically produces probes against all nodes of the clustering tree, providing an extensive coverage of known variants from a conserved functional gene. Using this programme, Rinta-Kanto and colleagues (2011) developed a taxon-specific microarray targeting sulfur-related gene transcription in members of *Roseobacter* clade, using data from 13 genome sequences. This FGA consisted of 1578 probes to 431 genes and was applied to the study of diverse natural *Roseobacter* communities. The results revealed that dimethylsulfoniopropionate was not preferred over other organic carbon and sulfur substrates by these populations.

ProDesign, developed by Feng and Tillier (2007), uses similar clustering methods with the aim of detecting all members of a same gene family in environmental samples. But, unlike HPD, this software uses spaced seed hashing, rather than a suffix tree algorithm, in order to benefit from permitted mismatches between a probe and its targets, and ensures the re-clustering of groups for which no probe was found. This results in a significant improvement in sequence coverage. As with HPD, however, this tool does not provide probes targeting uncharacterized nucleic acid sequences. In addition, to the best of our knowledge, no application using this design strategy has been reported in literature.

Although both of these strategies allow a wider range of sequence variants to be covered, and, therefore, appear best suited to describe microbial communities from complex environments, their main drawbacks are their inability to generate explorative probes and the absence of specificity tests (i.e. searching for potential cross-hybridizations) against large databases representative of microbial diversity. Recently, an efficient functional microarray probe design algorithm, called HiSpOD (High Specific Oligo Design), was proposed to overcome this problem (Dugat-Bony *et al.*, 2011). It is particularly useful for studying microbial communities in their environmental context. HiSpOD takes into account classical parameters for the design of effective probes (probe

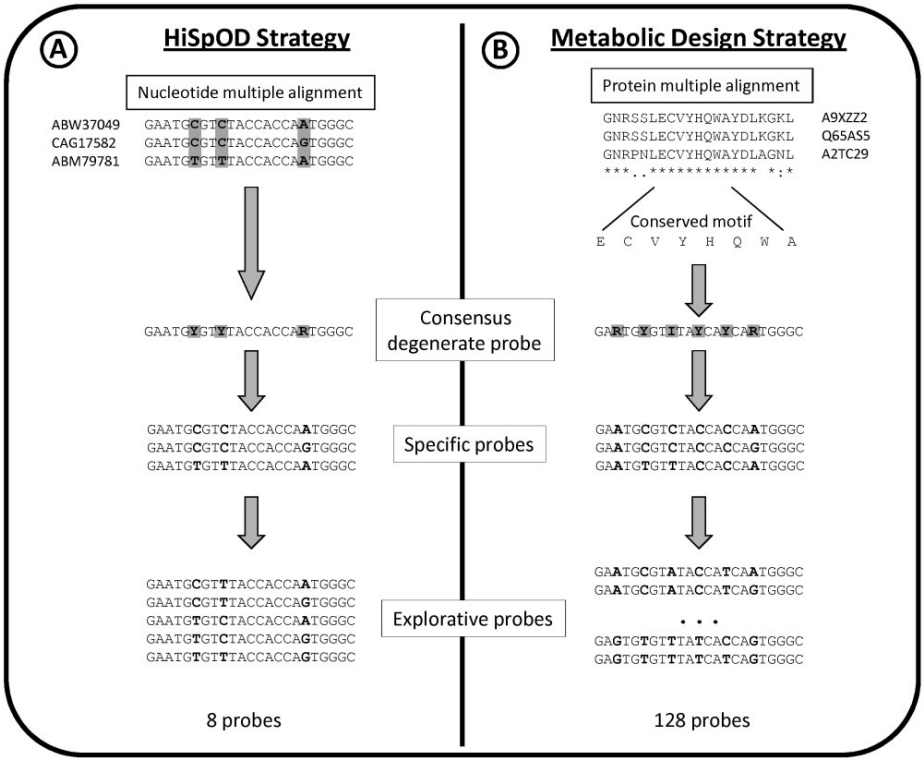


Fig. 2. Explorative probe design strategies implemented in (A) HiSpOD and (B) Metabolic Design software. The example shows probe design for the *bphA1c* gene encoding the Salicylate 1-hydroxylase alpha subunit involved in PAH degradation from three distinct *Sphingomonas* or *Sphingobium* species with both strategies.

length, Tm, GC%, complexity) and combines supplemental properties not considered by previous programmes. First, it can allow for the design of degenerate probes for gene families after multiple alignments of nucleic sequences belonging to the same gene family, and the production of consensus sequences. All combinations deduced from these degenerate probes are then divided into two groups. The first corresponds to specific probes for sequences available in databanks, and the second to explorative probes, which represent potential new signatures not corresponding to any previously described microorganisms (Fig. 2A). Both the probe sets covering the most likely gene sequence variants and those covering new combinations not yet deposited in databanks are created based on multiple mutation events already identified. Second, the specificity of all selected probes is checked against a large formatted database dedicated to microbial communities, the EnvExBase (Environmental Expressed sequences dataBase) composed of all coding DNA sequences (CDSs) from Prokaryotes (PRO), Fungi (FUN) and Environmental (ENV) taxonomic divisions of the EMBL databank, in order to limit cross-hybridizations. To validate this strategy, a microarray focusing on the genes involved in chloroethene solvent biodegradation was developed as a model system and enabled the identification of active cooperation between *Sulfurospirillum* and *Dehalococcoides* populations in the decontamination of a polluted groundwater (Dugat-Bony *et al*., 2011).

Use of protein sequence signatures for probe design

Unlike the strategies outlined above, a number of new strategies have been proposed to initiate probe design not from nucleic acid sequences, but from conserved peptidic regions, in order to survey all potential nucleic acid variants.

The first strategy based on this principle was described by Bontemps and colleagues (2005) and called CODEHMOP (for COnsensus DEgenerate Hybrid Motif Oligonucleotide Probe). It comes from an adaptation of the CODEHOP (for COnsensus DEgenerate Hybrid Oligonucleotide Primer) PCR primer design strategy, originally developed to identify distantly related genes encoding proteins that belong to known families (Rose *et al*., 1998; 2003; Boyce *et al*., 2009). In the CODEHMOP strategy, conserved amino acid motifs are identified from multiple alignments of protein sequences. Then, all possible

nucleic combinations (15–21 nucleotides) from the most highly conserved region (5–7 amino acids) of each protein motif are recreated and flanked by 5′ and 3′ fixed ends (12–15 nucleotides each), derived from the most frequent nucleotide at each position. The final probes are called 'hybrids', as they consist of a variable central core, to target a larger diversity, with some nucleic combinations not corresponding to any yet described sequences, and two fixed end sequences (available in databanks) added to increase probe length. The authors used this approach to design a prototype DNA array covering all described and undescribed *nodC* (nodulation gene) sequences in bacteria, and applied it to legume nodules (Bontemps *et al*., 2005). This strategy allowed the authors to detect new *nodC* sequences exhibiting less than 74% identity with known sequences.

The application of the CODEHMOP strategy is limited by the fact that it is not implemented into a fully automated programme and no probe specificity test is incorporated. Nevertheless, this approach appears to be the most comprehensive way of encompassing the larger diversity of gene sequence variants potentially found for enzymes mediating a given function. Furthermore, Terrat and colleagues (2010) developed a new software programme called Metabolic Design, which ensures *in silico* reconstruction of metabolic pathways, the identification of conserved motifs from protein multiple alignments and the generation of efficient explorative probes through a simple convenient graphical interface. In this case, before the probe design stage, the user reconstructs the chosen metabolic pathway *in silico* with all substrates and products from each metabolic step. One reference enzyme for each of these steps is selected and its protein sequence extracted from a curated database (by default, Swiss-Prot), which is then used to retrieve all homologous proteins from complete databases (Swiss-Prot and TrEMBL). After selecting the most pertinent homologous sequences, they are aligned to begin the probe design stage. The amino acids are back-translated for each molecular site identified, taking into account all genetic code redundancy, to produce a degenerate nucleic consensus sequence. All degenerate probes that meet the criteria defined by the user are retained (probe length and maximal degeneracy). All the specific possible combinations for each degenerate probe are subsequently checked for potential cross-hybridizations against a representative database (i.e. EnvExBase as in the HiSpOD programme). Finally, an output file, listing all degenerate probes selected by the user, permits the deduction of all possible combinations and organizes them into specific probes and exploratory probes (Fig. 2B). The approach was validated by studying enzymes involved in the degradation of polycyclic aromatic hydrocarbons (Terrat *et al*., 2010).

Towards circumventing microarray limitations

Despite the emergence of new design strategies, such as those presented above, the determination of a high-quality probe set appears to be crucial, especially in an environmental ecology context (Liebich *et al*., 2006; Leparc *et al*., 2009). Although explorative potential represents a major criterion for fingerprint determination, other parameters also impact considerably on probe sensitivity and specificity, and, therefore, require particular attention (Zhou, 2003; Wagner *et al*., 2007).

Optimization of probe size criterion

Generally, POAs employed for microbial community analysis contain short probes (typically 24–25 mers) (Brodie *et al*., 2006; Paliy *et al*., 2009; Rajilic-Stojanovic *et al*., 2009), whereas FGAs are built either with short (15–30 mers) (Bodrossy *et al*., 2003; Stralis-Pavese *et al*., 2004) or long oligonucleotides (40–70 mers) (Kane *et al*., 2000; Relogio *et al*., 2002; He *et al*., 2007). The main limitation of microarrays based on short oligonucleotide probes, therefore, is the need to use, in most cases, PCR-amplified targets to ensure enrichment and thereby increase sensitivity, but this also introduces an inherent PCR bias (Suzuki and Giovannoni, 1996; Peplies *et al*., 2004; Vora *et al*., 2004).

An alternative approach to design oligonucleotide probes, which combines excellent specificity with a potentially high sensitivity, is the use of the GoArrays strategy developed by Rimour and colleagues (2005) (software available at http://g2im.u-clermont1.fr/serimour/goarrays.html). In this approach, the oligonucleotide probe consists of the concatenation of two short subsequences that are complementary to disjoined regions of the target, with an insertion of a short random linker (e.g. 3–6 mer) (Fig. 3). This strategy has been shown to improve microarray efficiency for a wide range of applications (Rimour *et al*., 2005; Zhou *et al*., 2007; Pariset *et al*., 2009; Kang *et al*., 2010).

Specificity improvement using large databases

Because only a small portion of the natural microbial diversity has been identified, it is a major challenge to design appropriate probes specific to unique markers that do not cross-hybridize with similar unknown sequences (Chandler and Jarrell, 2005). Most of the currently available probe design software have been developed for non-environmental applications and performs specificity tests only against a reduced set of sequences, such as whole-genome data or specific sets of genes (Lemoine *et al*., 2009). The study of microbial communities, however, requires dedicated databases that are as representative

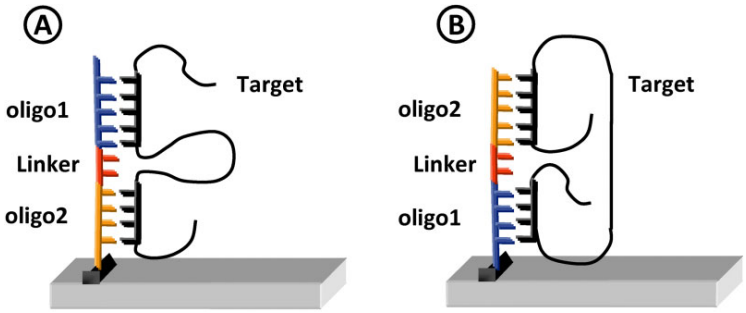


Fig. 3. Representation of the GoArrays strategy. In this strategy, two short oligonucleotide probes are concatenated with a random linker. Depending on the probes' positions, the target can form two kinds of stable loops during hybridization (A and B). (For the color version of this figure, please refer to the Web version of this article.)

as possible of all non-target sequences potentially present in environmental samples. GenBank (Benson *et al*., 2011), European Nucleotide Archive (ENA) (Leinonen *et al*., 2011) and the DNA Data Bank of Japan (DDBJ) (Kaminuma *et al*., 2011) are the most complete nucleic sequence databases publicly available to perform specificity tests. Dealing with such databases, however, is too time-consuming for probe design task, and, in this instance not really appropriate as some subsets of these databases correspond to sequences from organisms such as *Metazoa*, which are typically not considered in microbial ecology. Furthermore, for studies focusing on particular biomarkers, other sequence information need not to be considered.

For example, within POAs, each probe must be specific with respect to all small subunit (SSU) rRNA sequences, which may be present in the sample during hybridization. Curated and dedicated secondary databases have been already constructed [RDP (Cole *et al*., 2009), Greengenes (DeSantis *et al*., 2006) and SILVA (Pruesse *et al*., 2007)], assembling all SSU rRNA sequences described on public databases. The differences between these databases come from the construction and update pipelines that lead to distinct sizes: SILVA (Release 104) contains 1 304 069 16S rRNA sequences, RDP (Release 10) 1 545 680 and Greengenes (03/22/2011) 855 446. These large databases, therefore, are well adapted to phylogenetic probe design. PhylArray software (Militon *et al*., 2007) was developed before these databases were publicly available, and, therefore, uses its own highly curated (full length and quality filtered) and automatically updated prokaryotic SSU rRNA database (122 337 sequences for the last release).

Because environmental FGAs target coding sequences (CDS), the database used for specificity tests must include all known CDSs that may be encountered in natural environments. To the best of our knowledge, EnvExBase (integrated in both HiSpOD and Metabolic Design programmes) is the first CDSs database dedicated to microbial ecology (Terrat *et al*., 2010; Dugat-Bony *et al*., 2011). For its construction, all annotated transcript sequences and their associated 5′ and 3′ untranslated regions in all classes of EMBL Prokaryotes (PRO), Fungi (FUN) and Environmental (ENV) taxonomic divisions, were extracted and curated to remove bad-quality sequences. It represents a 9 129 323 sequence database.

The rapid growth of datasets, particularly environmental datasets, has led to an important increase in computational requirements coupled with a fundamental change in the way algorithms are conceived and designed [e.g. mpiBLAST (Darling *et al*., 2003)]. Consequently, parallel computing is essential, and algorithms must be deployed on large cluster infrastructures or computing grids, if specificity tests and alignments are to be performed with reasonable data processing times (Gardner *et al*., 2006; Thorsen *et al*., 2007).

Adaptation of the microarray format to the design strategy

Explorative design strategies targeting unknown sequences involve the use of degenerate probes (Bontemps *et al*., 2005; Militon *et al*., 2007; Terrat *et al*., 2010; Dugat-Bony *et al*., 2011). Consequently, the selected strategy will greatly influence the choice between the two major DNA microarray types (*ex situ* or *in situ*), the platform and the density (Dufva, 2005; Ehrenreich, 2006; Kawasaki, 2006). When using *in situ* synthesis microarrays, such as the Agilent, Affymetrix and NimbleGen platforms, all combinations resulting from a degenerate probe must be independently synthesized. This will exponentially increase the final number of probes for the array production (density). For instance, concerning the CODEHMOP (Bontemps *et al*., 2005) and Metabolic Design strategies (Terrat *et al*., 2010), because the genetic code often involves degeneracy at the third position of each codon, a 24 mer probe (targeting a seven amino acid conserved motif) will generate at least 128 combinations (assuming a minimal degeneracy rate of two for each codon). This value will reach at least 131 072 for a 51 mer probe containing 17 degenerate positions. Conversely, *ex situ* platforms allow the degenerate probes (all combinations mixed together) to be spotted in the same location

Table 3. Characteristics of the main commercially distributed high-density microarrays.

Type of array	Technology	Probe length	Max features	Max plex
Spotted Arrays	Robot spotting Pre-made DNA	Any	~100 000	1
Affymetrix	Photolithography *in situ*	<100 mer (generally 25 mer)	~6 000 000	1
NimbleGen	Micro-mirrors *in situ*	<100 mer (generally 50–75 mer)	4 200 000	12
Agilent	Inkjet *in situ*	<100 mer (generally 25–60 mer)	1 000 000	8

on the array and consequently reduce the total amount of features.

Other user choices may also affect the final number of probes per array. Replication is crucial to achieve reliable data for microarrays (Spruill *et al*., 2002). Multiple replicates of the same probe provide some back-up in case a feature cannot be evaluated due to technical artefacts, such as dye precipitations or dust particles. A statistical estimation has deduced that at least three replicates should be made (Lee *et al*., 2000). Second, multiple probes per gene could be designed in order to increase confidence in the results (Loy *et al*., 2002; Chou *et al*., 2004) and to mask misleading signal variations whose causes (e.g. target secondary structure, probe folding, etc.) are not yet fully understood (Pozhitkov *et al*., 2007). Third, some platforms, such as Affymetrix GeneChips, determine probe pairs where each probe ('match') is accompanied by a negative control with a single differing base in the middle of the probe ('mismatch probe') in order to discriminate between real signals and those due to non-specific hybridizations (Lipshutz *et al*., 1999).

To address this problem of probe number, several commercial companies have proposed two major types of high-density microarrays whose main characteristics are described in Table 3: (i) *in situ* synthesized microarrays, distributed by Agilent (http://www.chem.agilent.com), NimbleGen (http://www.nimblegen.com) and Affymetrix (http://www.affymetrix.com), which can attain billions of probes and be physically divided into multi-arrays per slide (up to 12) to perform simultaneous analyses of several samples on a single experiment; and (ii) spotted microarrays [e.g. Arrayit (http://www.arrayit.com)] with a current printing capacity close to 100 000 features per microarray.

Concluding remarks and future directions

Assessing the extreme microbial diversity encountered in environmental samples represents an exciting challenge that could create a better understanding of microbial community functioning. Environmental DNA microarrays, with the opportunity to survey both known and unknown microorganisms through explorative probe design, are one of the most powerful approaches for achieving this goal. Future perspectives in this domain will be to systematically integrate this innovative concept into probe design workflows, especially by offering the possibility to design degenerate probes targeting sequence clusters. Furthermore, to efficiently recognize signals due to unknown targets, it will be particularly useful to develop automatic procedures to analyse microarray data. In addition, using explorative probe design in sequence capture approaches that couple with NGS, such as those originally developed for direct selection of human genomic loci (Albert *et al*., 2007), could also improve this gene characterization. Indeed, sequence capture elution products should allow the full identification and characterization of new taxa when using POAs or new protein coding genes with FGAs.

The constant increase in available sequences (Cochrane *et al*., 2009) means that databases for specificity tests must be regularly updated. As a result, probe datasets must be re-computed as frequently as possible in order to take into account all deposited data. Nevertheless, assessing probe specificity against large databases is a time-consuming task. To overcome this problem, two complementary strategies could be employed: (i) Creation of databases specific to each ecological compartment. Usually, specificity tests are not performed against a suitable subset of sequences mainly due to lack of databases for microbial ecology. Depending on the environment studied it would be more relevant to perform these tests against reduced databanks dedicated to specific ecosystems (soil, marine, freshwater, gut, etc.).

(ii) Parallelization of probe design algorithms. Perspectives to limit computation time are based on exploiting the computational resources available using specialized frameworks such as Message Passing Interface (MPI) or heterogeneous systems including General-purpose Processing on Graphics Processing Units (GPGPU). With the recent development of extremely fast broadband networks, it has become possible to distribute the calculations at larger and larger scales over different geographical locations (Schadt *et al*., 2010). Cluster, grid or emerging cloud computing are all examples of shared computing resources where probe design algorithms can be deployed. Being able to improve the bioinformatics tools applied to environmental microbiology through algorithm deployment on such shared computational resources, and combining them with automatic update pipelines, are two important challenges and strategies for the future of the field of molecular ecology.

Acknowledgments

This work was supported by the grant ID 2598 from the 'Agence De l'Environnement et de la Maîtrise de l'Energie' (ADEME, France); the Grant ANR-07-ECOT-005–05 for the programme PRECODD Evasol from 'Agence Nationale de la Recherche' (ANR, France); the Grant ANR-08-BIOENERGIES-0 for the programme BIOENERGIES AnaBio-H2 from 'Agence Nationale de la Recherche' (ANR, France); and the INSU-EC2CO programme from 'Centre National de la Recherche Scientifique' (CNRS, France). We thank David Tottey for reviewing the English version of the manuscript.

References

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4:** 903–905.

Amann, R., Ludwig, W., and Schleifer, K. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143–169.

Ashelford, K.E., Weightman, A.J., and Fry, J.C. (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res* **30:** 3481–3489.

Bae, J.W., and Park, Y.H. (2006) Homogeneous versus heterogeneous probes for microbial ecological microarrays. *Trends Biotechnol* **24:** 318–323.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2011) GenBank. *Nucleic Acids Res* **39:** D32–D37.

Bers, K., Sniegowski, K., Albers, P., Breugelmans, P., Hendrickx, L., De Mot, R., and Springael, D. (2011) A molecular toolbox to estimate the number and diversity of Variovorax in the environment: application in soils treated with the phenylurea herbicide linuron. *FEMS Microbiol Ecol* **76:** 14–25.

Blaskovic, D., and Barak, I. (2005) Oligo-chip based detection of tick-borne bacteria. *FEMS Microbiol Lett* **243:** 473–478.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environ Microbiol* **5:** 566–582.

Boeckaert, C., Vlaeminck, B., Fievez, V., Maignien, L., Dijkstra, J., and Boon, N. (2008) Accumulation of trans C18:1 fatty acids in the rumen after dietary algal supplementation is associated with changes in the Butyrivibrio community. *Appl Environ Microbiol* **74:** 6923–6930.

van den Bogert, B., de Vos, W.M., Zoetendal, E.G., and Kleerebezem, M. (2011) Microarray Analysis and Barcoded Pyrosequencing Provide Consistent Microbial Profiles Depending on the Source of Human Intestinal Samples. *Appl Environ Microbiol* **77:** 2071–2080.

Bommarito, S., Peyret, N., and SantaLucia, J., Jr (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res* **28:** 1929–1934.

Bontemps, C., Golfier, G., Gris-Liebe, C., Carrere, S., Talini, L., and Boivin-Masson, C. (2005) Microarray-based detection and typing of the Rhizobium nodulation gene nodC: potential of DNA arrays to diagnose biological functions of interest. *Appl Environ Microbiol* **71:** 8042–8048.

Boyce, R., Chilana, P., and Rose, T.M. (2009) iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. *Nucleic Acids Res* **37:** W222–W228.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B., and DeRisi, J.L. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol* **4:** R9.

Brodie, E.L., Desantis, T.Z., Joyner, D.C., Baek, S.M., Larsen, J.T., Andersen, G.L., *et al.* (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl Environ Microbiol* **72:** 6288–6298.

Brodie, E.L., DeSantis, T.Z., Parker, J.P., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci USA* **104:** 299–304.

Candela, M., Consolandi, C., Severgnini, M., Biagi, E., Castiglioni, B., Vitali, B., *et al.* (2010) High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction – universal array approach. *BMC Microbiol* **10:** 116.

Castiglioni, B., Rizzi, E., Frosini, A., Sivonen, K., Rajaniemi, P., Rantala, A., *et al.* (2004) Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. *Appl Environ Microbiol* **70:** 7161–7172.

Chandler, D.P., and Jarrell, A.E. (2005) Taking arrays from the lab to the field: trying to make sense of the unknown. *Biotechniques* **38:** 591–600.

Chou, C.C., Chen, C.H., Lee, T.T., and Peck, K. (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res* **32:** e99.

Chung, W.-H., Rhee, S.-K., Wan, X.-F., Bae, J.-W., Quan, Z.-X., and Park, Y.-H. (2005) Design of long oligonucleotide probes for functional gene detection in a microbial community. *Bioinformatics* **21:** 4092–4100.

Claesson, M.J., O'Sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J.R., Smidt, H., *et al.* (2009) Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. *PLoS ONE* **4:** e6669.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., *et al.* (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res* **37:** D19–D25.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

Cruz-Martinez, K., Suttle, K.B., Brodie, E.L., Power, M.E.,

Andersen, G.L., and Banfield, J.F. (2009) Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J* **3:** 738–744.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99:** 10494–10499.

Darling, A., Carey, L., and Feng, W. (2003) The Design, Implementation, and Evaluation of mpiBLAST. *In: 4th International Conference on Linux Clusters: The HPC Revolution 2003. San Jose, California*.

DeAngelis, K.M., Brodie, E.L., DeSantis, T.Z., Andersen, G.L., Lindow, S.E., and Firestone, M.K. (2009) Selective progressive response of soil microbial community to wild oat roots. *ISME J* **3:** 168–178.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., *et al*. (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* **77:** 1315–1324.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al*. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* **53:** 371–383.

Duc, L., Neuenschwander, S., Rehrauer, H., Wagner, U., Sobek, J., Schlapbach, R., and Zeyer, J. (2009) Development and experimental validation of a nifH oligonucleotide microarray to study diazotrophic communities in a glacier forefield. *Environ Microbiol* **11:** 2179–2189.

Dufva, M. (2005) Fabrication of high quality microarrays. *Biomol Eng* **22:** 173–184.

Dugat-Bony, E., Missaoui, M., Peyretaillade, E., Biderre-Petit, C., Bouzid, O., Gouinaud, C., *et al*. (2011) HiSpOD: probe design for functional DNA microarrays. *Bioinformatics* **27:** 641–648.

Ehrenreich, A. (2006) DNA microarray technology for the microbiologist: an overview. *Appl Microbiol Biotechnol* **73:** 255–273.

Emrich, S.J., Lowe, M., and Delcher, A.L. (2003) PROBEmer: A web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Res* **31:** 3746–3750.

Evertsz, E.M., Au-Young, J., Ruvolo, M.V., Lim, A.C., and Reynolds, M.A. (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* **31:** 1182. 1184, 1186 passim.

Feldhaar, H., Straka, J., Krischke, M., Berthold, K., Stoll, S., Mueller, M.J., and Gross, R. (2007) Nutritional upgrading for omnivorous carpenter ants by the endosymbiont Blochmannia. *BMC Biol* **5:** 48.

Feng, S., and Tillier, E.R.M. (2007) A fast and flexible approach to oligonucleotide probe design for genomes and gene families. *Bioinformatics* **23:** 1195–1202.

Flanagan, J.L., Brodie, E.L., Weng, L., Lynch, S.V., Garcia, O., Brown, R., *et al*. (2007) Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *J Clin Microbiol* **45:** 1954–1962.

Franke-Whittle, I.H., Goberna, M., Pfister, V., and Insam, H. (2009) Design and development of the ANAEROCHIP microarray for investigation of methanogenic communities. *J Microbiol Methods* **79:** 279–288.

Fraune, S., Augustin, R., Anton-Erxleben, F., Wittlieb, J., Gelhaus, C., Klimovich, V.B., *et al*. (2010) In an early branching metazoan, bacterial colonization of the embryo is controlled by maternal antimicrobial peptides. *Proc Natl Acad Sci USA* **107:** 18067–18072.

Gans, J., Wolinsky, M., and Dunbar, J. (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309:** 1387–1390.

Gardner, M.K., Feng, W.-C., Archuleta, J., Lin, H., and Ma, X. (2006) Parallel genomic sequence-searching on an ad-hoc grid: experiences, lessons learned, and implications. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. SC Conference (ed.). Tampa, Florida: ACM, pp. 22.

Gentry, T., Wickham, G., Schadt, C., He, Z., and Zhou, J. (2006) Microarray Applications in Microbial Ecology Research. *Microb Ecol* **52:** 159–175.

Gittel, A., Sorensen, K.B., Skovhus, T.L., Ingvorsen, K., and Schramm, A. (2009) Prokaryotic community structure and sulfate reducer activity in water from high-temperature oil reservoirs with and without nitrate treatment. *Appl Environ Microbiol* **75:** 7086–7096.

Guerrero, R., and Berlanga, M. (2006) Life's unity and flexibility: the ecological link. *Int Microbiol* **9:** 225–235.

He, Z., Gentry, T.J., Schadt, C.W., Wu, L., Liebich, J., Chong, S.C., *et al*. (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1:** 67–77.

He, Z., Deng, Y., Van Nostrand, J.D., Tu, Q., Xu, M., Hemme, C.L., *et al*. (2010) GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *ISME J* **4:** 1167–1179.

He, Z., Van Nostrand, J.D., Deng, Y., and Zhou, J.Z. (2011) Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. *Front Environ Sci Engin China* **5:** 1–20.

He, Z.L., Van Nostrand, J.D., Wu, L.Y., and Zhou, J.Z. (2008) Development and application of functional gene arrays for microbial community analysis. *Trans Nonferrous Met Soc China* **18:** 1319–1327.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19:** 342–347.

Huyghe, A., Francois, P., Charbonnier, Y., Tangomo-Bento, M., Bonetti, E.-J., Paster, B.J., *et al*. (2008) Novel Microarray Design Strategy To Study Complex Bacterial Communities. *Appl Environ Microbiol* **74:** 1876–1885.

Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., *et al*. (2011) DDBJ progress report. *Nucleic Acids Res* **39:** D22–D27.

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D., and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28:** 4552–4557.

Kang, S., Denman, S.E., Morrison, M., Yu, Z., Dore, J., Leclerc, M., and McSweeney, C.S. (2010) Dysbiosis of

fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. *Inflamm Bowel Dis* **16:** 2034–2042.

Kawasaki, E.S. (2006) The end of the microarray Tower of Babel: will universal standards lead the way? *J Biomol Tech* **17:** 200–206.

Kelly, J.J., Siripong, S., McCormack, J., Janus, L.R., Urakawa, H., El Fantroussi, S., *et al*. (2005) DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. *Water Res* **39:** 3229–3238.

Kelly, L.C., Cockell, C.S., Piceno, Y.M., Andersen, G.L., Thorsteinsson, T., and Marteinsson, V. (2010) Bacterial diversity of weathered terrestrial Icelandic volcanic glasses. *Microb Ecol* **60:** 740–752.

Klitgaard, K., Boye, M., Capion, N., and Jensen, T.K. (2008) Evidence of multiple Treponema phylotypes involved in bovine digital dermatitis as shown by 16S rRNA gene analysis and fluorescence in situ hybridization. *J Clin Microbiol* **46:** 3012–3020.

Koltai, H., and Weingarten-Baror, C. (2008) Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic Acids Res* **36:** 2395–2405.

Kyselkova, M., Kopecky, J., Felfoldi, T., Cermak, L., Omelka, M., Grundmann, G.L., *et al*. (2008) Development of a 16S rRNA gene-based prototype microarray for the detection of selected actinomycetes genera. *Antonie Van Leeuwenhoek* **94:** 439–453.

Lee, M.L., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA* **97:** 9834–9839.

Lehner, A., Loy, A., Behr, T., Gaenge, H., Ludwig, W., Wagner, M., and Schleifer, K.H. (2005) Oligonucleotide microarray for identification of Enterococcus species. *FEMS Microbiol Lett* **246:** 133–142.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., *et al*. (2011) The European Nucleotide Archive. *Nucleic Acids Res* **39:** D28–D31.

Lemoine, S., Combes, F., and Le Crom, S. (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Res* **37:** 1726–1739.

Leparc, G.G., Tuchler, T., Striedner, G., Bayer, K., Sykacek, P., Hofacker, I.L., and Kreil, D.P. (2009) Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Res* **37:** e18.

Letowski, J., Brousseau, R., and Masson, L. (2004) Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. *J Microbiol Methods* **57:** 269–278.

Li, F., and Stormo, G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **17:** 1067–1076.

Li, X., He, Z., and Zhou, J. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res* **33:** 6114–6123.

Liebich, J., Schadt, C.W., Chong, S.C., He, Z., Rhee, S.K., and Zhou, J. (2006) Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol* **72:** 1688–1691.

Liles, M.R., Turkmen, O., Manske, B.F., Zhang, M., Rouillard, J.-M., George, I., *et al*. (2010) A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. *Soil Biol Biochem* **42:** 739–747.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat Genet* **21:** 20–24.

Loy, A., and Bodrossy, L. (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin Chim Acta* **363:** 106–119.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., *et al*. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68:** 5064–5081.

Loy, A., Schulz, C., Lucker, S., Schopfer-Wendels, A., Stoecker, K., Baranyi, C., *et al*. (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order 'Rhodocyclales'. *Appl Environ Microbiol* **71:** 1373–1386.

Ludwig, W., Amann, R., Martinez-Romero, E., Schönhuber, W., Bauer, S., Neef, A., and Schleifer, K.-H. (1998) rRNA based identification and detection systems for rhizobia and other bacteria. *Plant Soil* **204:** 1–19.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Kumar, Y., *et al*. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32:** 1363–1371.

Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Preheim, S.P., *et al*. (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proc Natl Acad Sci USA* **103:** 13629–13634.

Militon, C., Rimour, S., Missaoui, M., Biderre, C., Barra, V., Hill, D., *et al*. (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics* **23:** 2550–2557.

Muhling, M., Woolven-Allen, J., Murrell, J.C., and Joint, I. (2008) Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. *ISME J* **2:** 379–392.

Neufeld, J.D., Mohn, W.W., and de Lorenzo, V. (2006) Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. *Environ Microbiol* **8:** 126–140.

Nordberg, E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics* **21:** 1365–1370.

Øvreås, L. (2000) Population and community level approaches for analysing microbial diversity in natural environments. *Ecol Lett* **3:** 236–251.

Pace, N.R. (1997) A Molecular View of Microbial Diversity and the Biosphere. *Science* **276:** 734–740.

Paliy, O., Kenche, H., Abernathy, F., and Michail, S. (2009) High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray. *Appl Environ Microbiol* **75:** 3572–3579.

Palmer, C., Bik, E.M., Eisen, M.B., Eckburg, P.B., Sana, T.R., Wolber, P.K., *et al*. (2006) Rapid quantitative profiling

of complex microbial populations. *Nucleic Acids Res* **34:** e5.

Pariset, L., Chillemi, G., Bongiorni, S., Romano Spica, V., and Valentini, A. (2009) Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *Nat Biotechnol* **25:** 272–279.

Peplies, J., Lau, S.C.K., Pernthaler, J., Amann, R., and Glöckner, F.O. (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environ Microbiol* **6:** 638–645.

Pozhitkov, A., Noble, P.A., Domazet-Loso, T., Nolte, A.W., Sonnenberg, R., Staehler, P., *et al.* (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res* **34:** e66.

Pozhitkov, A.E., Tautz, D., and Noble, P.A. (2007) Oligonucleotide microarrays: widely applied – poorly understood. *Brief Funct Genomic Proteomic* **6:** 141–148.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188–7196.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008) The rational exploration of microbial diversity. *ISME J* **2:** 997–1006.

Rajilic-Stojanovic, M., Heilig, H.G., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., and de Vos, W.M. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol* **11:** 1736–1751.

Rastogi, G., Osman, S., Vaishampayan, P.A., Andersen, G.L., Stetler, L.D., and Sani, R.K. (2010a) Microbial diversity in uranium mining-impacted soils as revealed by high-density 16S microarray and clone library. *Microb Ecol* **59:** 94–108.

Rastogi, G., Osman, S., Kukkadapu, R., Engelhard, M., Vaishampayan, P.A., Andersen, G.L., and Sani, R.K. (2010b) Microbial and mineralogical characterizations of soils collected from the deep biosphere of the former Homestake gold mine, South Dakota. *Microb Ecol* **60:** 539–550.

Rastogi, G., Barua, S., Sani, R.K., and Peyton, B.M. (2011) Investigation of Microbial Populations in the Extremely Metal-Contaminated Coeur d'Alene River Sediments. *Microb Ecol* **62:** 1–13.

Relogio, A., Schwager, C., Richter, A., Ansorge, W., and Valcarcel, J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res* **30:** e51.

Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G., and Fayard, J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* **20:** 271–273.

Rhee, S.K., Liu, X., Wu, L., Chong, S.C., Wan, X., and Zhou, J. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl Environ Microbiol* **70:** 4303–4317.

Rimour, S., Hill, D., Militon, C., and Peyret, P. (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* **21:** 1094–1103.

Rinta-Kanto, J.M., Burgmann, H., Gifford, S.M., Sun, S., Sharma, S., del Valle, D.A., *et al.* (2011) Analysis of sulfur-related transcription by Roseobacter communities using a taxon-specific functional gene microarray. *Environ Microbiol* **13:** 453–467.

Roh, S.W., Abell, G.C.J., Kim, K.-H., Nam, Y.-D., and Bae, J.-W. (2010) Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends Biotechnol* **28:** 291–299.

Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., and Henikoff, S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* **26:** 1628–1635.

Rose, T.M., Henikoff, J.G., and Henikoff, S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res* **31:** 3763–3766.

Rouillard, J.M., Zuker, M., and Gulari, E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* **31:** 3057–3062.

Rusch, A., and Amend, J.P. (2004) Order-specific 16S rRNA-targeted oligonucleotide probes for (hyper)thermophilic archaea and bacteria. *Extremophiles* **8:** 357–366.

Sagaram, U.S., DeAngelis, K.M., Trivedi, P., Andersen, G.L., Lu, S.E., and Wang, N. (2009) Bacterial diversity analysis of Huanglongbing pathogen-infected citrus, using PhyloChip arrays and 16S rRNA gene clone library sequencing. *Appl Environ Microbiol* **75:** 1566–1574.

Sanguin, H., Sarniguet, A., Gazengel, K., Moenne-Loccoz, Y., and Grundmann, G.L. (2009) Rhizosphere bacterial communities associated with disease suppressiveness stages of take-all decline in wheat monoculture. *New Phytol* **184:** 694–707.

SantaLucia, J., Jr, Allawi, H.T., and Seneviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35:** 3555–3562.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., and Nolan, G.P. (2010) Computational solutions to large-scale data management and analysis. *Nat Rev Genet* **11:** 647–657.

Schliep, A., and Rahmann, S. (2006) Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree. *Bioinformatics* **22:** e424–e430.

Schonmann, S., Loy, A., Wimmersberger, C., Sobek, J., Aquino, C., Vandamme, P., *et al.* (2009) 16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genus Burkholderia. *Environ Microbiol* **11:** 779–800.

Severgnini, M., Cremonesi, P., Consolandi, C., Caredda, G., De Bellis, G., and Castiglioni, B. (2009) ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. *Nucleic Acids Res* **37:** e109.

Shawkey, M.D., Firestone, M.K., Brodie, E.L., and Beissinger, S.R. (2009) Avian incubation inhibits growth and diversification of bacterial assemblages on eggs. *PLoS ONE* **4:** e4522.

Simon, C., and Daniel, R. (2009) Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* **85:** 265–276.

Spruill, S.E., Lu, J., Hardy, S., and Weir, B. (2002) Assessing sources of variability in microarray gene expression data. *Biotechniques* **33:** 916–920. 922–913.

Stralis-Pavese, N., Sessitsch, A., Weilharter, A., Reichenauer, T., Riesing, J., Csontos, J., *et al*. (2004) Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. *Environ Microbiol* **6:** 347–363.

Suzuki, M.T., and Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625–630.

Taroncher-Oldenburg, G., Griner, E.M., Francis, C.A., and Ward, B.B. (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl Environ Microbiol* **69:** 1159–1171.

Terrat, S., Peyretaillade, E., Goncalves, O., Dugat-Bony, E., Gravelat, F., Mone, A., *et al*. (2010) Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics* **11:** 478.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

Thorsen, O., Smith, B., Sosa, C.P., Jiang, K., Lin, H., Peters, A., and Feng, W.-C. (2007) Parallel genomic sequence-search on a massively parallel system. In *Proceedings of the 4th International Conference on Computing Frontiers*. ACM (ed.). Ischia, Italy: ACM, pp. 59–68.

Tiquia, S.M., Wu, L., Chong, S.C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques* **36:** 664–670.

Torsvik, V., and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* **5:** 240–245.

Vora, G.J., Meador, C.E., Stenger, D.A., and Andreadis, J.D. (2004) Nucleic acid amplification strategies for DNA microarray-based pathogen detection. *Appl Environ Microbiol* **70:** 3047–3054.

Wagner, M., Smidt, H., Loy, A., and Zhou, J. (2007) Unravelling Microbial Communities with DNA-Microarrays: Challenges and Future Directions. *Microb Ecol* **53:** 498–506.

Wang, X., and Seed, B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* **19:** 796–802.

Weinert, N., Piceno, Y., Ding, G.C., Meincke, R., Heuer, H., Berg, G., *et al*. (2011) PhyloChip hybridization uncovered an enormous bacterial diversity in the rhizosphere of different potato cultivars: many common and few cultivar-dependent taxa. *FEMS Microbiol Ecol* **75:** 497–506.

Wernersson, R., and Nielsen, H.B. (2005) OligoWiz 2.0 – integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res* **33:** W611–W615.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95:** 6578–6583.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002) High-Density Microarray of Small-Subunit Ribosomal DNA Probes. *Appl Environ Microbiol* **68:** 2535–2541.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* **87:** 4576–4579.

Xu, J. (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol* **15:** 1713–1731.

Yergeau, E., Schoondermark-Stolk, S.A., Brodie, E.L., Dejean, S., DeSantis, T.Z., Goncalves, O., *et al*. (2009) Environmental microarray analyses of Antarctic soil microbial communities. *ISME J* **3:** 340–351.

Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005) Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol Bioeng* **89:** 670–679.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **6:** 288–294.

Zhou, Z., Dou, Z.-X., Zhang, C., Yu, H.-Q., Liu, Y.-J., Zhang, C.-Z., and Cao, Y.-J. (2007) A strategy to optimize the oligo-probes for microarray-based detection of viruses. *Virol Sin* **22:** 326–335.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.

Pouvoir explorer l'immense biodiversité présente dans l'environnement est encore aujourd'hui l'un des enjeux majeurs de l'écologie microbienne. Les biopuces ADN possèdent désormais les atouts nécessaires pour atteindre cet objectif notamment grâce i) aux nouveaux concepts de *design* de sondes dites « exploratoires » qui permettent d'accéder à la fraction inconnue des communautés microbiennes, ii) à la puissance des outils informatiques permettant de s'assurer de la pertinence des sondes sélectionnées et iii) au développement de nouveaux formats de biopuces ADN pouvant contenir jusqu'à un million de sondes.

Alors que les approches de séquençage massif comme la métagénomique permettent d'explorer sans *a priori* les écosystèmes d'intérêt et d'identifier de nouvelles espèces ou de nouveaux gènes, d'autres stratégies sont basées sur le principe inverse. La diversité n'est plus recherchée directement dans l'écosystème d'intérêt, elle est, en effet, imaginée au travers des signatures nucléiques dégénérées prenant en compte l'ensemble de la variabilité des séquences nucléiques ou protéiques déjà disponibles dans les bases de données. Ainsi, à partir des signatures identifiées, de nouvelles combinaisons de séquences sont créées, séquences valides au niveau génétique et donc potentiellement portées par des individus présents dans les écosystèmes, bien que jamais identifiées auparavant. Ce concept, que l'on peut qualifier d'« exploratoire », a d'abord été appliqué pour la détermination d'amorces PCR et validé par la caractérisation de nouvelles séquences parfois très éloignées de celles déjà connues (Rose *et al.* 1998 ; 2003). Son utilisation s'étend maintenant à d'autres outils de biologie moléculaire comme les biopuces ADN (Bontemps *et al.* 2005) ou la capture de gènes (Denonfoux *et al.* 2013). Cependant, elle nécessite le développement d'outils informatiques adaptés.

3.2 Outils logiciels pour la sélection de sondes oligonucléotidiques

L'évolution constante des puissances de calcul et des infrastructures informatiques, de même que le nombre de séquences disponibles dans les bases de données, offrent de nouvelles perspectives pour les stratégies de détermination de sondes oligonucléotidiques dédiées aux problématiques d'écologie microbienne.

De manière à proposer une vue d'ensemble des stratégies bioinformatiques disponibles pour la détermination de sondes oligonucléotidiques appliquée aux biopuces ADN, la rédaction d'un chapitre d'un ouvrage intitulé « *Microarrays: Current Technology, Innovations and Applications* » (http://www.horizonpress.com/microarrays2), a été réalisé sous la direction du Dr. Zhili He de l'Université d'Oklahoma (USA).

Chapitre d'ouvrage n°1

Software Tools for the Selection of Oligonucleotide Probes for Microarrays.

In, Microarrays: Current Technology, Innovations and Applications.

Software Tools for the Selection of Oligonucleotide Probes for Microarrays

2

Nicolas Parisot, Jérémie Denonfoux, Eric Dugat-Bony, Eric Peyretaillade and Pierre Peyret

Abstract

Oligonucleotide microarrays have been widely used for gene detection and quantification of gene expression. Recently, they have been adapted for profiling microbial communities in a flexible and easy-to-use manner. In fact, it is possible to analyse both the microbial diversity and the metabolic capacity of complex communities in one experiment. However, the quality of the result is largely dependent on the quality of designed probes. Probe design, which is not a trivial task, should thus take into account multiple parameters such as the oligonucleotide sequence and its binding capacity in order to ensure high specificity, sensitivity, and uniformity as well as potentially quantitative capability for each probe. Furthermore, the exploration of the not-yet-described fraction of complex communities requires consideration of the explorative power of oligonucleotide probes. To design such probes, multiple tools have been developed based on different algorithms. These algorithms and the different probe criteria that they used are described in the present chapter. However, the best algorithm to guarantee a high-quality design must be chosen with the knowledge of biological questions and biological samples.

Introduction

With exponential growth in the availability of complete genome sequences and metagenomic data sets and the low cost of DNA synthesis, oligonucleotide arrays have become the most widely used type of microarrays. Furthermore, with the advancement of microarray technology (e.g. *in situ* synthesis technologies), high-density oligonucleotide microarrays can hold millions of probes on a single microscopic glass slide with multiplexing capacities. These molecular tools can be easily synthesized on demand, in small batches, and at low cost. This flexibility combined with rapid data acquisition, management and interpretation allow oligonucleotide microarrays to continue to advance next-generation sequencing in various applications. Several strategies using oligonucleotides probes have been developed to improve the specificity and sensitivity of gene detection (Fig. 2.1). The most widely used strategy (Fig. 2.1A) is based on the determination of specific subsequences in the targeted genes that serve as probes. The subsequent steps involve the hybridization of the labelled targets, followed by the image processing. Other supplementary steps are added to improve the specificity and sensitivity of detection.

The capture of targets is strongly influenced by their secondary and tertiary structures; therefore, probes should be directed towards accessible regions. However, measuring or predicting the effect of secondary structure is still difficult. The shearing of target molecules into small fragments is one widely utilized technique. Alternatively, to overcome secondary structure constraints, a two-probe proximal chaperone detection system (Fig. 2.1B) that consists of a species-specific capture probe and a chaperone probe (sometimes also used as a labelled detector) that reduces secondary structure formation was developed (Small *et al.*, 2001). The term 'chaperone probe' has been used rather than the term 'stacking probe' (Yershov *et al.*, 1996), which was originally used in the

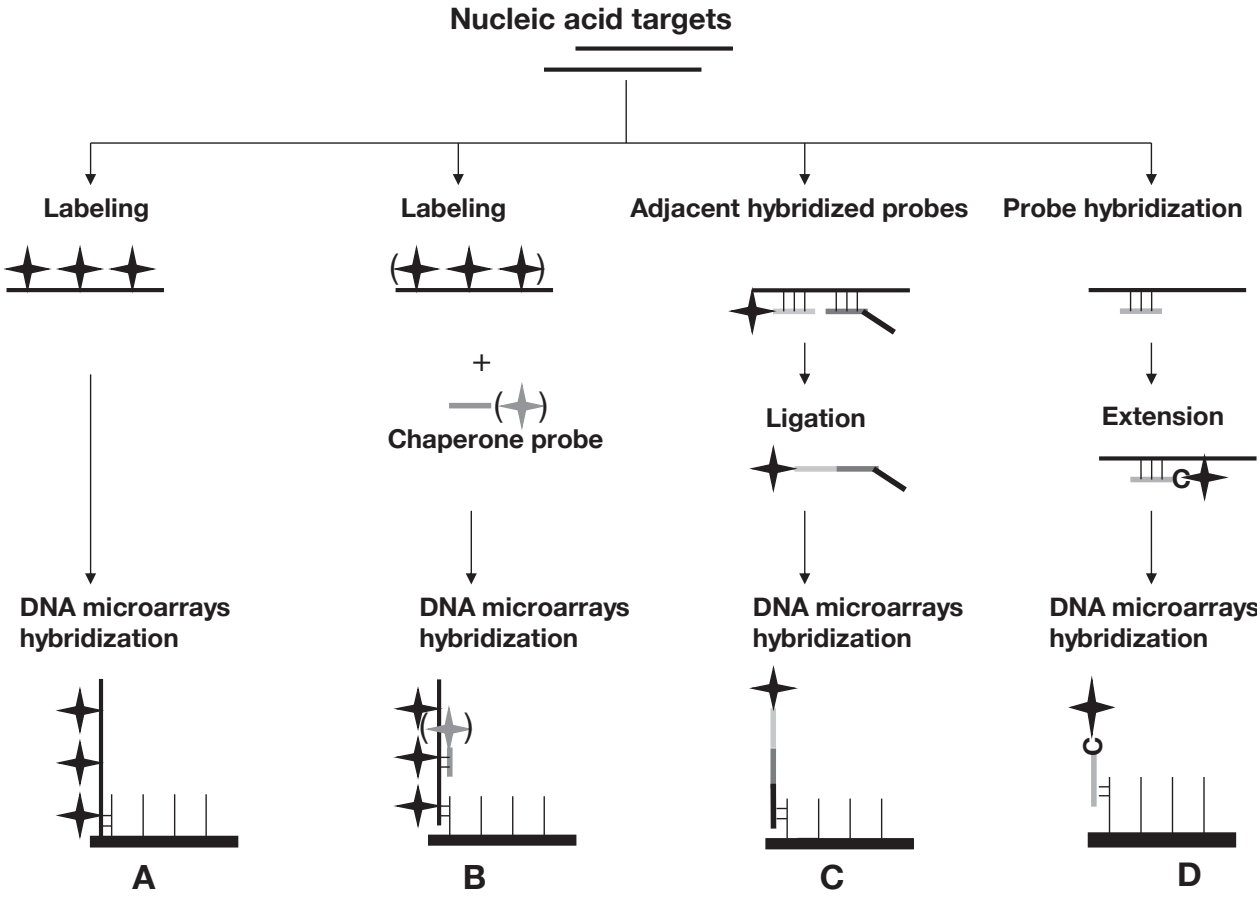


Figure 2.1 Schematic representation of different oligonucleotide-based approaches used in DNA microarray technology. (A) The most widely used approach allows the direct recognition of labelled targets by specific probes fixed on a solid surface. (B) The two-probe proximal chaperone detection system uses helper probes (chaperones) to resolve the secondary or tertiary structure of the targets, thereby improving accessibility for efficient matching between the specific probe located on the solid surface of the DNA microarrays and the target. Labelling could be directed towards the targets (black) or the chaperones (grey). (C) Enzymatic ligation uses a high-selectivity ligase, which requires the perfect complementarity of the double-stranded DNA structure to successfully catalyse the covalent joining of two adjacently hybridized probes. Detection is only possible if the probes are linked together. In the universal microarray approach, the common probe has a tag sequence (cZip code in black) that directs the hybridization on the capture probe (Zip) on the solid surface of the DNA microarrays. (D) The single nucleotide extension labelling allows the reverse complement probe to be labelled in a linear amplification reaction given the availability of the corresponding target sequence. The labelled reverse complement probe is then captured by specific probes located on the solid surface of the DNA microarrays.

context of polymorphism detection (Maldonado-Rodriguez *et al.*, 1999). However, chaperone detector probes located in the immediate proximity of the capture probe provide detectable, non-specific, non-target binding, presumably because of base-stacking effects (Chandler *et al.*, 2003). In some cases, the addition of specific DNA helper oligonucleotides improved detection (Kaplinski *et al.*, 2010). However, the use of helper oligonucleotides is not always practical because of the difficulty of designing helper probes with the same specificity as the capture probe but without non-target detection (Peplies *et al.*, 2003).

Enzymatic ligation (Fig. 2.1C) is another microarray-based method that has also been used for the detection of environmental microorganisms (Gerry *et al.*, 1999; Busti *et al.*, 2002; Castiglioni *et al.*, 2004; Hultman *et al.*, 2008; Candela *et al.*, 2010). The reaction is performed separately from array hybridization, which enables the use of address (also known as tag or zip) oligonucleotides to equalize probe hybridization conditions. The enzymatic ligation step is the

primary source of specificity. The principle of detecting specific DNA templates by enzymatic ligation was developed to overcome some of the limitations of oligomeric hybridization probes in distinguishing single base mutations associated with genetic diseases. Enzymatic ligation relies on the high selectivity of the ligase, which requires the perfect complementarity of a double-stranded DNA structure to successfully catalyse the covalent joining of two adjacently hybridized probes. The probes constitute a target-specific probe pair that becomes detectable only if the probes are linked together. The so-called discriminating probe is designed such that the 3′ end matches the target at a unique position containing a nucleotide that distinguishes the target from other species. A common probe is designed to hybridize adjacent to the discriminating probe, which enables ligation if an appropriate target is present in the reaction mixture. In the universal microarray approach (Gerry *et al.*, 1999), the common probe has a 3′-tag sequence (cZip code) that directs it to the correct address on the array, whereas the discriminating probe is fluorescently labelled.

The advantages of the universal array lie in the uniform hybridization conditions of all zip sequences and in flexibility, as the same array platform can be used with multiple ligation probe sets. ORMA (oligonucleotide retrieving for molecular applications) is a set of scripts for searching discriminating positions and selecting oligonucleotide probes for such an approach (ligase detection reaction; LDR) or for Minisequencing/Primer Extension (Severgnini *et al.*, 2009). A variant strategy that utilizes a cleavable padlock probe has recently been developed (van Doorn *et al.*, 2009) that eliminates the probe amplification of the initial padlock probe assay (Szemes *et al.*, 2005), resulting in a background-free assay. Padlock probes are long oligonucleotides that contain asymmetric target complementary regions at both their 5′ and 3′ ends to confer specific target detection. Upon hybridization to the target, the two ends are brought into contact, which allows probe circularization by ligation. In the first assay after exonuclease treatment, the circularized probes are amplified and hybridized on DNA microarrays. The central part of the probe harbours sequences for PCR amplification and DNA microarray capture. In a recent improvement to the method, in addition to the sequence complementary to the probe on DNA microarrays, padlock probes now harbour a cleavage site in their central part near the labelling position. After cleavage, only the originally ligated padlock probes can be visualized on the DNA microarray.

Finally, a microbial diagnostic microarray approach using single nucleotide extension labelling (SSELO: sequence-specific end labelling of oligonucleotides) has been developed (Kostić *et al.*, 2007). Reverse complements of the capture oligonucleotides (RC oligonucleotides) are end-labelled in a linear amplification reaction based on the availability of the corresponding target sequence (Fig. 2.1D). The entire mixture is hybridized to the microarray to identify the sequences that have been labelled. The specificity of the assay was shown to be determined primarily by the stringency of the annealing step during labelling rather than that of the subsequent hybridization.

Regardless of the strategy used to develop DNA microarrays, probe selection remains the key element in obtaining an efficient detection tool. Several criteria related to probe characteristics influence the efficiency of detection and should be assessed with caution before fabricating DNA microarrays.

General criteria for probe design

Specificity is defined according to the ability of the probe to not cross-hybridize with non-target sequences (i.e. probes should discriminate well between the intended target and all other sequences present in the target pool). Sensitivity is defined as the strength with which a probe binds to its target. This parameter influences the level of the detection signal, and consequently, the relevance of obtained information (i.e. probes should detect differences in target concentrations under given hybridization conditions). Uniformity corresponds to the similarity of hybridization behaviour for a given probe set, i.e. similar thermodynamic characteristics under the same experimental conditions (e.g. temperature, salt and formamide concentration), which could also

influence sensitivity to some extent. In fact, the structural properties of several probes, including probe length, GC content, melting temperature (T_m) and the Gibbs free energy reflecting binding capacities (ΔG), are optimized in this case.

Specificity

The ability to minimize or eliminate cross-hybridization is an important parameter and represents a current bottleneck in the design of microarray probes (Wernersson *et al.*, 2007). In fact, the specificity of the hybridization of a probe with its target is one of the most important parameters that determines the quality of the microarray result (Kane *et al.*, 2000; Koltai and Weingarten-Baror, 2008). Specificity is defined as the ability of a probe to bind to a target sequence without hybridization to non-targets. Currently, most probe design software uses the BLAST algorithm (Altschul *et al.*, 1990) to search for potential cross-hybridization against custom databases constructed in concordance with microarray experiments and applications. Probe specificity assessment with BLAST uses a homology threshold that determines whether the oligonucleotide is specific. Kane's recommendations for long oligonucleotides are based on the discarding of probes that share a total identity greater than 75–80% or contiguous stretches of identity greater than 15 nucleotides with a non-target sequence (Kane *et al.*, 2000). Alternatively, some probe design software uses a suffix array approach to overcome BLAST's limitations (Manber and Myers, 1993). Rather than performing several local alignments, the suffix array method utilizes an efficient and space-saving data structure that quickly identifies and records, in alphabetical order, all possible substrings or suffixes and their locations in the input sequences. The theory of suffix arrays states that the longest common prefix (LCP) shared by any two non-adjacent suffixes must be equal to or shorter than the LCP of any two neighbouring suffixes between them in the suffix array (Manber and Myers, 1993; Chou *et al.*, 2004a). The main limitation of this approach is the memory storage of the suffix structure. For example, the human genome, which has three billion characters, requires 12 GB for storage of the entire suffix array (Sadakane and Shibuya, 2001). Thermodynamic calculations are also used to evaluate the strength of cross-hybridizations by determining the binding-free energy between the probe and the non-target sequence to give an indication of the duplex's stability. As the probe is bound to a solid surface rather than being free in solution, the calculation appears as an approximation (Pozhitkov *et al.*, 2007). Finally, other probe design software uses custom methodologies to evaluate probe specificity based, for instance, on global alignments or hierarchical clustering approaches (Lemoine *et al.*, 2009). Even with the use of BLAST or suffix array tools combined with thermodynamic prediction, several other criteria must be considered during the design process to improve probe specificity.

Low-complexity regions such as those containing long homopolymers may also contribute to probe specificity and consequently must be avoided during the probe design process (Wang and Seed, 2003; Leparc *et al.*, 2009). To overcome this problem, many probe design algorithms, such as CommOligo (Li *et al.*, 2005), ROSO (Reymond *et al.*, 2004) or HiSpOD (Dugat-Bony *et al.*, 2011), apply a filter or mask these particular nucleotide repeats, whereas YODA (Nordberg, 2005) can discard specific regions defined beforehand as prohibited for the probe design. These particular regions can also be highlighted by more complex calculations using a lossless compression algorithm such as the LZW compression algorithm (Ziv and Lempel, 1977), a suffix array structure or custom calculations for complexity scoring. Otherwise, low-complexity regions can be masked using the DUST program (Hancock and Armstrong, 1994) included in the software that uses the BLAST algorithm for the assessment of potential cross-hybridization.

Just as it can enhance probe sensitivity, the probe's position on the sequence could also influence the oligonucleotide specificity (Tomiuk and Hofmann, 2001). For example, the 3′ untranslated region (3′UTR) in eukaryotic mRNA is considered the less-conserved region because of the usage of alternative polyadenylation signals (Tomiuk and Hofmann, 2001). Consequently, the choice of 3′UTRs for probe design reduces

the probability of cross-hybridization with closely related paralogues. However, the potential alternative polyadenylation signals found in 3′UTR combined with a propensity for repetitive elements has to be taken in consideration. Thus, some programmes compute a localization score based on the distance to the centre or to the 3′ or 5′ end of the sequence [e.g. OligoWiz (Wernersson and Nielsen, 2005)] or let the user localize the designed probes in a 3′ or 5′ range [e.g. OligoPicker (Wang and Seed, 2003)]. Others can display [e.g. YODA (Nordberg, 2005)] all of the non-overlapping probes or only those located in the 3′ end, 5′ end or in the centre of the sequence.

Sensitivity

The term sensitivity is closely related to the affinity of a probe to its target, which is mediated by hybridization and is characterized by the free energy difference ΔG that measures the binding affinity for the two strands to form a duplex. ΔG can be estimated from the probe sequence using nearest neighbour models that provide a reasonable approximation of ΔG for strands hybridizing in solution (SantaLucia, 1998). Furthermore, in microarray experiments where quantitative detection is required, microarray probes should also exhibit a sensitive and predictable response to concentrations of specific targets (Mei *et al.*, 2003). Although in-solution parameters, for example, base composition, temperature and salt concentration, are typically used for such calculations, the estimated *Tm* of the nucleotide duplex is a good proxy for the sensitivity of the probe to some extent. Nevertheless, even though the thermodynamic properties of nucleic acid duplex formation and dissociation in solution are well known (SantaLucia, 1998), the thermodynamic properties during hybridization at the solid–liquid interface in a microarray context remain unclear (Pozhitkov *et al.*, 2007). Thus, several parameters have to be considered to increase probe sensitivity and allow microarray probes to exhibit a sensitive and predictable response to a target concentration.

Although the secondary structure must be considered as the main sensitivity criterion for microarray design, the probe length, number of probes per target and probe position can also be considered. The choice of which criteria to use will typically depend on the probe design strategy and microarray synthesis technology.

Secondary structure

To achieve maximum probe sensitivity, the design must exclude oligonucleotides that are able to form homo-dimers or stable intramolecular secondary structures such as hairpins or stem–loops that may impact hybridization efficiency by preventing stable target hybridization (Lemoine *et al.*, 2009). Thus, the objective is to prevent the formation of any such structures at the hybridization temperature by assessing the secondary structure. Some probe design software uses alignment-based strategies for a self-annealing assessment combined with scoring calculations (Kämpke *et al.*, 2001), thermodynamic calculations based on the Mfold tool (Zuker, 2003) or suffix array data (Manber and Myers, 1993) to assess secondary structure stability in combination with a specificity test to evaluate potential cross-hybridization.

Probe length

Probe sensitivity generally increases with probe length, as the binding energy for longer probe–target duplexes is typically higher and hybridization kinetics are irreversible (Hughes *et al.*, 2001; Relógio *et al.*, 2002; Letowski *et al.*, 2004; Dugat-Bony *et al.*, 2012b). Long oligonucleotide probes (50- to 60-mers) have a comparable sensitivity to PCR-based probes with a length of 300–400 nucleotides. Fifty-mer probes demonstrate good specificity as long as the similarity with non-targeted sequences is less than 75% or there is no stretch of 15 perfectly matching nucleotides (Kane *et al.*, 2000). The use of 60-mer oligonucleotide probes for hybridization could allow the detection of targets with 8-fold higher sensitivity than the use of 25-mer probes (Chou *et al.*, 2004a), whereas an identity of less than 77% between a 60-mer probe and its target results in a lack of signal (Hughes *et al.*, 2001). Generally, the threshold for differentiation between targets is 75–90% (Kane *et al.*, 2000; Taroncher-Oldenburg *et al.*, 2003; Tiquia *et al.*, 2004; Dugat-Bony *et al.*, 2012b) identity for such long probes, which

indicates low specificity (Li *et al.*, 2005). In contrast, short oligonucleotide (18- to 30-mer) probes are more specific, as they allow the discrimination of single nucleotide polymorphisms under optimal hybridization conditions but with reduced sensitivity (Relógio *et al.*, 2002). The GoArrays strategy (Rimour *et al.*, 2005) combines both advantages by designing long probes (high sensitivity) composed of two short subsequences (high specificity).

Number of probes per target

As noted above, longer oligonucleotides provide higher sensitivity than shorter probes (Hughes *et al.*, 2001; Relógio *et al.*, 2002; Dugat-Bony *et al.*, 2012b). However, the use of one probe per gene with long oligonucleotide microarrays appears limiting even though oligonucleotide hybridization is highly sequence-dependent (Tijssen, 1993; Chou *et al.*, 2004a). In fact, the binding of an oligonucleotide probe to different regions of the target yields different signal intensities (Selinger *et al.*, 2000; Hughes *et al.*, 2001) and thus complicates the prediction of whether an oligonucleotide probe will bind efficiently to its target and yield a good hybridization signal based on sequence information alone (Chou *et al.*, 2004a). Thus, multiple probes per gene have been used in oligonucleotide array designs to obtain reliable quantitative information for gene expression (Selinger *et al.*, 2000; Hughes *et al.*, 2001) as well as gene detection in complex environmental samples (Dugat-Bony *et al.*, 2011, 2012a). Five probes per gene has been suggested as a suitable number for 30-mer probes (Relógio *et al.*, 2002), but this number could increase with probe length. A perfect case would be to select a minimal probe set that ensures good hybridization signals and test it experimentally to successfully detect targets even at low levels, but such a large-scale screening process remains extremely time-consuming and costly.

Probe position

The positioning of the probe along the target may also impact the hybridization signal, especially in gene expression experiments (Wernersson *et al.*, 2007). The signal may decrease near the 5′ end when using poly-T-primed cDNA synthesis, which requires a multiple-probe design along the length of the transcripts. The decreased signal is a consequence of the stability of the RNA (Auer *et al.*, 2003) and enzymatic reactions during sample preparation such as reverse transcription that have a tendency to terminate early (Wernersson *et al.*, 2007). Thus, the probes used for gene expression are preferentially positioned near the 3′ end of eukaryotic transcripts. In contrast, for cDNA synthesis using random priming in prokaryote gene expression experiments, decreases in signal can be observed for probes positioned at the very end of the 3′ end of the gene (Wernersson *et al.*, 2007).

Uniformity

Microarray technology relies on the simultaneous hybridization of many probes under the same conditions (e.g. salt concentrations, temperature); therefore, uniform thermodynamic behaviour for the selected probes is crucial (Loy and Bodrossy, 2006; Wagner *et al.*, 2007; Dugat-Bony *et al.*, 2012b). The easiest way to reach this objective is to select probes with homogeneous probe lengths, but several other parameters must be considered, such as the melting temperature (*Tm*) and GC content.

Tm uniformity

To achieve maximum homogeneity in the probe set, a primary objective is to select probes that share similar melting temperatures (T_m), which ensures quantitative comparison of gene expression and detection as well as similar microarray hybridization for all genes targeted in the study. Chemical compounds such as tetra-alkyl ammonium salts (Jacobs *et al.*, 1988), which have been applied in dot-blot experiments with degenerate oligonucleotide hybridization probes (Wood *et al.*, 1985), are known to eliminate the dependence of T_m on base composition. However, these salts have not been widely applied in microarray experiments. Thus, an alternative is to select oligonucleotide probes with melting temperatures that fall within a narrow range. Several methods are available to calculate the T_m of a probe; the most frequently used is the application of the nearest-neighbour (NN) model using parameters from SantaLucia (1998) or from Rychlik *et al.* (1990).

The T_m can be calculated directly by the probe design software, by an external program or by using a custom method. Most probe design software, for example CommOligo (Li *et al.*, 2005) and HiSpOD (Dugat-Bony *et al.*, 2011), allow the user to select a T_m range in which the selected probes will be designed. Some, such as OligoArray (Rouillard *et al.*, 2003), OligoPicker (Wang and Seed, 2003), PICKY (Chou *et al.*, 2004b) and YODA (Nordberg, 2005), perform optimization calculations by adapting some parameters, such as the probe length, to select probes in the expected T_m range. For some programmes, such as ArrayOligoSelector (Bozdech *et al.*, 2003) and ProbeSelect (Li and Stormo, 2001), $T_m$is not considered and selection is based solely on the similar T_m values of probes with uniform lengths and GC content. Finally, all of the available formulas calculate the T_m for oligonucleotides that are free in solution, and not oligonucleotide probes bound to a glass surface. However, the probe's behaviour in solution could be different from that when attached to a slide, and thus, it is more suitable that probes fall into a T_m range rather than having a precise T_m value.

GC content

The oligonucleotide GC content is another parameter to consider and is closely related to the melting temperature (T_m). Some probe design software, such as OligoWiz (Wernersson and Nielsen, 2005), OligoPicker (Wang and Seed, 2003) or ProbeSelect (Li and Stormo, 2001), does not consider the GC content as a potential criterion for the oligonucleotide probe selection process. In contrast, other software such as CommOligo (Li *et al.*, 2005), OligoArray (Rouillard *et al.*, 2003) or YODA (Nordberg, 2005) allow the user to select a GC content range and filter candidate probes that do not fulfil this range from the final probe list. Generally, the programmes use a preferential range between 40% and 65% (Lemoine *et al.*, 2009), and some are able to perform a T_m optimization by using the GC content range defined by the user to select the best probe candidates. This strategy appears useful for oligonucleotide probe design that involves sequences with very high or very low GC content.

Probe design algorithms for microbial DNA microarrays

Phylogenetic oligonucleotide arrays (POAs)

To rapidly characterize the members of microbial communities present in complex environments, numerous phylogenetic oligonucleotide arrays (POAs) have been developed using the SSU rRNA biomarker (Loy *et al.*, 2002; Wilson *et al.*, 2002; Brodie *et al.*, 2006, 2007; Palmer *et al.*, 2006; DeSantis *et al.*, 2007; Hazen *et al.*, 2010). Fully automated software and manual approaches have both been developed to design POAs (Tables 2.1–2.3).

Alignment-based strategies

Initially, probe design software for POAs was primarily based on aligned sequence sets such as PRIMROSE (Ashelford *et al.*, 2002), PROBE (Pozhitkov and Tautz, 2002), ARB-Probe Design (Ludwig *et al.*, 2004), PhylArray (Militon *et al.*, 2007) and ORMA (Severgnini *et al.*, 2009). Probe design software programmes based on aligned input data or on performing a multiple sequence alignment as the first step of the algorithm is well suited for the design of probes with an optimal coverage of the target group. Multiple sequence alignments are generally converted into consensus sequences that account for the sequence variability at each position. Then, probe design programs search for conserved regions to select oligonucleotides.

The Probe Design tool included in the ARB program package (Ludwig *et al.*, 2004) has been widely used to develop low-density, custom-made POAs for reduced groups of organisms (Loy *et al.*, 2005; Neufeld *et al.*, 2006; Franke-Whittle *et al.*, 2009). The ARB Probe Design is able to design oligonucleotides with a length of 10–100 nucleotides using a three-step algorithm. First, the user selects the target group through the ARB interface. Second, the program searches for potential target sites (avoiding repetitive regions) and subsequently returns a ranked list of candidate oligonucleotides according to several compositional and thermodynamic criteria. Finally, the proposed oligonucleotide probes are evaluated against the entire database using the Probe Match

Table 2.1 Comparison of probe design software features for phylogenetic oligonucleotide arrays (POAs): applications and availability

Software	Reference	Application	Availability	URL
ARB (v 5.3)	Ludwig *et al*. (2004)	POA	Downloadable, standalone GUI (L, M)	http://www.arb-home.de/
CaSSiS (v 0.5.0)	Bader *et al*. (2011)	POA	Downloadable, command-line (L)	http://cassis.in.tum.de
PhylArray	Militon *et al*. (2007)	POA	Web Interface	http://g2im.u-clermont1.fr/serimour/phylarray
ORMA	Severgnini *et al*. (2009)	POA, FGA	Matlab Script	Upon request
KASpOD	Parisot *et al*. (2012)	POA, FGA, WGA-ORF	Web interface or command-line (L)	http://g2im.u-clermont1.fr/kaspod/

POA, phylogenetic oligonucleotide array; FGA, functional gene array; WGA-ORF, open reading-frame oriented whole-genome array; GUI, graphical user interface. L, Linux; M, MacOS.

Table 2.2 Comparison of probe design software features for phylogenetic oligonucleotide arrays (POAs): main features

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
ARB (v 5.3)	Fixed by the user (10–100)	No localization specified	All probes reaching selection criteria	No	No	Yes	Yes	No	No
CaSSiS (v 0.5.0)	Fixed by the user or range chosen by the user	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	No	Yes	Yes	No	No
PhylArray	Fixed by the user (20–70)	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	No	Yes	Yes	No	Yes
ORMA	Fixed by the user	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	Yes	No	Yes	No	Yes
KASpOD	Fixed by the user (18–31)	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	No	No	No	No	Yes

tool. Local alignments are determined between the probe and the most similar sequences in the database, and up to five mismatches are allowed.

Among the alignment-based oligonucleotide design software programs for POA, ORMA (Oligonucleotide Retrieving for Molecular Applications) appears to be suitable for the determination of discriminating positions within a set of highly similar sequences (Severgnini *et al.*, 2009). This software is well adapted for the ligation or extension strategies described in the introduction to this chapter. ORMA relies on a Single Base Seeker (SBS) algorithm to locate positions that are able to discriminate one sequence from a set of closely related sequences. First, the user selects the sequences that are to

Table 2.3 Comparison of probe design software features for phylogenetic oligonucleotide arrays (POAs): flexibility

Software	Organism	Cross-hybridization assessment	Database for specificity test	Input files
ARB (v 5.3)	No limitation	Local alignment and thermodynamic calculations	ARB-Silva database	ARB database. Nucleotide sequences.
CaSSiS (v 0.5.0)	No limitation	ARB Positional Tree server and distance calculations	Input sequence dataset	FASTA file with all target and non-target sequences and a list or a tree file containing targeted sequence identifiers. Nucleotide sequences.
PhylArray	Prokaryotes	BLAST and Kane's specifications	Custom non-redundant SSU rRNA database (95MB)	No input files are required.
ORMA	No limitation	No	No	Multiple sequence alignment file (Clustal-like, Multi Sequence Files, or aligned FASTA format). Nucleotide sequences.
KASpOD	No limitation	Global alignment and distance calculations	External FASTA file	Two FASTA files (targeted sequences and non-target sequences). Nucleotide sequences.

be considered as targeted from among the dataset; the remaining sequences are subsequently used as the group from which the discriminating positions must be different. Then, for each non-degenerate position, the SBS algorithm calculates the sum of sequences carrying the same base as the considered sequence. If the only sequence harbouring this base is the targeted sequence, the position is identified as discriminant. This last step is reiterated, replacing each degenerate position (except for undetermined and subsequently non-discriminant positions referred to as N's) with its two or three alternative bases. Candidate oligonucleotides are then defined at these discriminating positions by retrieving flanking sequences. A series of constraints and quality filters is used to assign a quality score to each putative probe (i.e. length, melting temperature, number of degenerate bases, low-complexity regions). Moreover, intra-group (i.e. coverage) and inter-group (i.e. specificity) scores are calculated, and the probes that maximize the intra-group score and have the lowest inter-group score are selected.

The design strategies described above are not solely dedicated to high-density microarrays (i.e. those with tens of thousands of oligonucleotide probes). High-density POAs are, however, the most promising approach to comprehensive screening all known bacterial and archaeal taxa with a single microarray (Dugat-Bony *et al.*, 2012b). Many strategies for designing large probe sets are not fully automated and thus not provided in the form of autonomous software. For instance, the PhyloChip (DeSantis *et al.*, 2007), which is the most widely used high-density POA, was constructed using a semi-automated procedure explained in the supplementary material of Hazen *et al.* (Hazen *et al.*, 2010). All known 16S rRNA sequences containing at least 1300 nucleotides were extracted from the NAST multiple sequence alignment (DeSantis *et al.*, 2006a) of the Greengenes (DeSantis *et al.*, 2006b) database. Then, sequences were filtered to remove putative chimeras using the Bellerophon software (Huber *et al.*, 2004) and also to remove low-complexity sequences (i.e. sequences with more than three homopolymers with a length greater or equal to eight) and sequences with ambiguous nucleotides (i.e. sequences with ambiguous base calls greater than or equal to 0.3%). Retained 16S rRNA sequences were then clustered at 0.5% sequence divergence in 59,959 operational taxonomic units (OTUs). The 59,959 OTUs represented 1464 families, 1219 orders, 1123 classes, 147 phyla and two domains. For each OTU, each of the sequences within the OTU was separated into overlapping 25-mers segments, and these potential targets were used to select the probe

set. Candidate 25-mer oligonucleotides were selected from the subalignment according to thermodynamic constraints (i.e. GC content, secondary structure, melting temperature, and self-dimerization). Potential targets were ranked according to their universality among the OTU; those having data for all members of the OTU were preferred over those found in only a fraction of the OTU members. Candidate probes that matched exactly with well-ranked putative targets were selected for microarray fabrication.

Computational alignment of a large multiple sequences is a time-consuming task. Thus, to accelerate the computations, the probe design software tools have to be retooled to permit the computation of many probes based on large sequence datasets.

Alignment-free strategies

CaSSiS (Comprehensive and Sensitive Signature Search) was developed to address the limited ability of previous probe design software to handle large collections of sequences (Bader *et al.*, 2011). CaSSiS is able to perform fast and comprehensive probe design based on a three-step algorithm. First, CaSSiS extracts and assesses each possible probe. The results are stored in a bipartite graph where the probes' coverage within the overall dataset is represented as edges (Fig. 2.2). Evaluating all of the probes could be a time-consuming task, but CaSSiS uses the ARB Positional Tree Server (PT-Server) (Ludwig *et al.*, 2004) to rapidly identify exact and inexact matches. Based on predefined parameters such as length or the number of mismatches allowed, the PT-Server, using a truncated suffix tree, returns all matches of the query probe. Because CaSSiS supports a relaxed search within the database, the user can specify the number of mismatches allowed within the targeted sequences and the mismatch threshold for non-target hits. The second stage of the CaSSiS algorithm consists of ranking candidate oligonucleotides according to their specificity scores. The last step extracts the probes that harbour the highest coverage and have up to *n* non-target matches (outgroup hits), where *n* is user-defined.

Even if the probe design software for POAs were able to handle millions of sequences, its capabilities would always be restricted to surveying

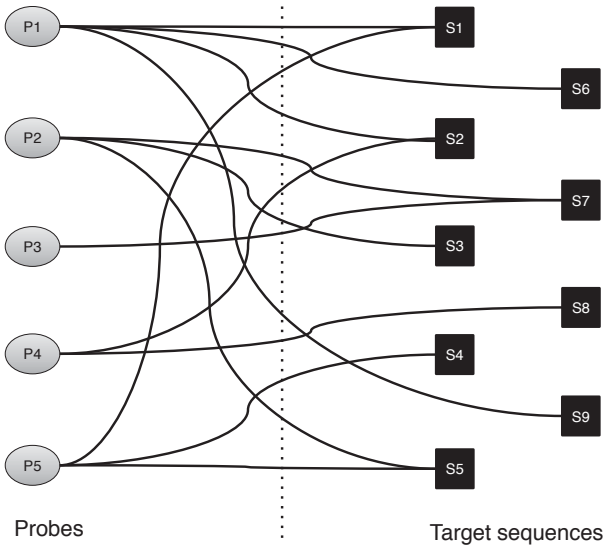


Figure 2.2 A Bipartite graph. This data structure provides a representation of the probes' coverage. Nodes represent a probe (P) or a sequence (S), and edges indicate which sequences are matched with which probe.

known microorganisms with sequences that have been deposited in a public database. However, in spite of the high number of recorded sequences, our current vision of microbial diversity is still incomplete, partially because of the tremendous diversity of microbial species, ecological niches and technological limits. Detection of 90% of the richness in some complex environments could require tens of thousands of times the current sequencing effort (Quince *et al.*, 2008). A major challenge, therefore, is to develop new strategies for designing explorative probes to target sequences that have not yet been described.

Explorative probe design strategies for POA

There are two ways to detect unknown microorganisms: using probes defined from known high phylogenetic levels and using explorative probes that correspond to new sequence variants of existing phylogenetic signatures that are not yet deposited in public databases but potentially present in the environment.

The 'multiple probe concept' consists of several probes to target an organism at different phylogenetic levels (e.g. genus, family, order). This strategy dramatically reduces the risk of misidentification and substantially increases the

resolution of the analysis by discriminating bacteria down to the species level (Loy and Bodrossy, 2006; Schliep and Rahmann, 2006; Huyghe *et al.*, 2008; Schönmann *et al.*, 2009; Liles *et al.*, 2010). The use of this strategy to construct POAs is well suited to ensuring the detection of unknown microorganisms by probes defined at higher taxonomic levels. Nevertheless, such probes are strictly complementary to known sequences and do not harbour the explorative power to detect microorganisms with uncharacterized phylogenetic signatures (Dugat-Bony *et al.*, 2012b).

The first software program dedicated to POAs that offered the possibility of designing explorative probes was the PhylArray program (Militon *et al.*, 2007). PhylArray was developed to survey whole microbial communities, including known and unknown microorganisms, in complex environments. The first step of the PhylArray algorithm (Fig. 2.3) is the extraction of all available sequences corresponding to a targeted taxon from a custom 16S rRNA curated database. Retrieved sequences are then aligned using the ClustalW program (Thompson *et al.*, 1994). A degenerate consensus sequence is then deduced from this multiple sequence alignment, taking into account the sequence variability at each position. Degenerate candidate probes are then selected along the consensus sequence, and all non-degenerate combinations are checked for cross-hybridizations against the 16S rRNA database. Among the combinations derived from each degenerate probe, some correspond to sequences that have not yet been deposited in public databases, namely explorative probes. Such probes should, therefore, allow the detection of undescribed microorganisms belonging to the targeted taxon. Probes defined using this software, which were recently used to evaluate the bacterial diversity in soils (Delmont *et al.*, 2011), yield a higher sensitivity and specificity than probes designed using the PRIMROSE and ARB strategies (Militon *et al.*, 2007). PhylArray was designed to account for all of the sequence variability within the targeted sequences, but because it relies on multiple sequence alignment, it is limited in its ability to manage large input datasets. Consequently, new probe design strategies are needed to define explorative probes based on large databases.

KASpOD (Parisot *et al.*, 2012) software was developed to overcome this limitation. KASpOD (K-mer Based Algorithm for Highly Specific and

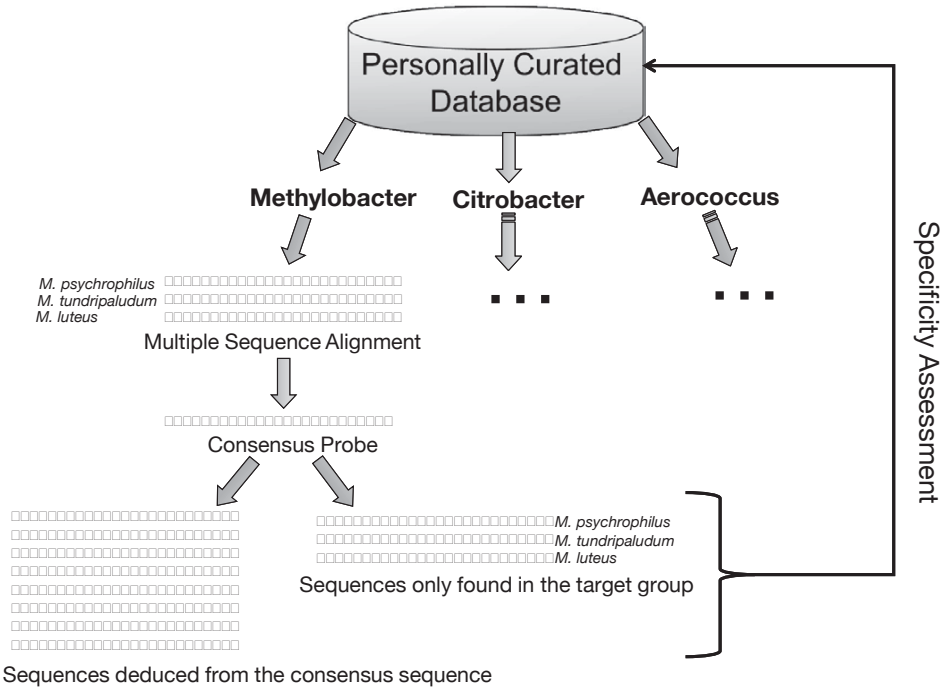


Figure 2.3 PhylArray program workflow. The PhylArray program is composed of four steps: (i) sequence extraction for each taxon, (ii) multiple sequence alignment, (iii) degenerate consensus sequence production and probe selection and (iv) specificity tests against the 16S rRNA database.

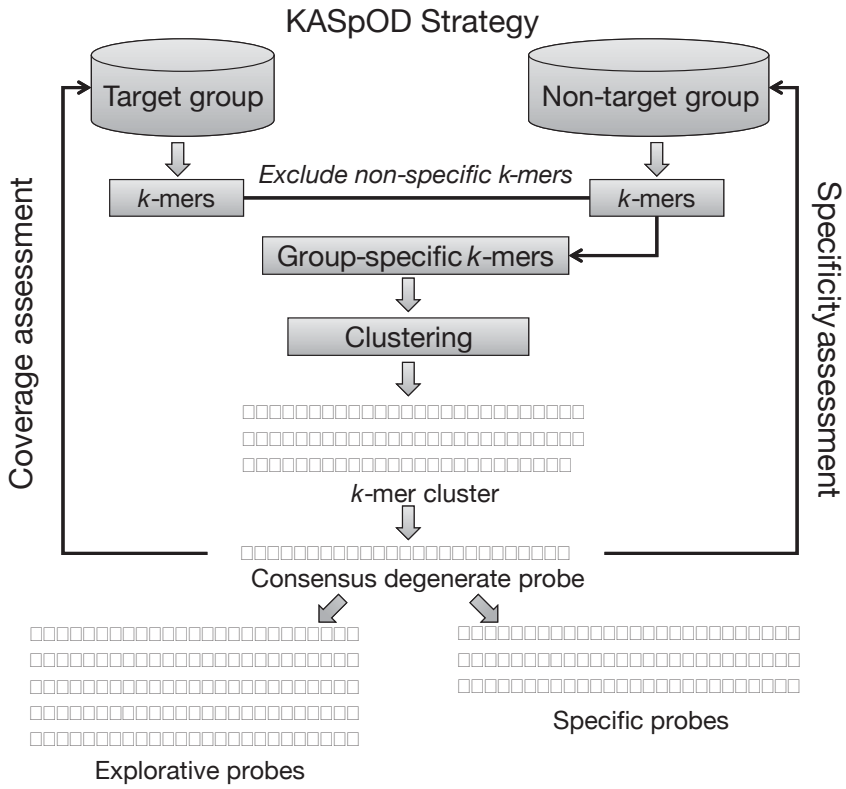


Figure 2.4 KASpOD program workflow. The KASpOD program is composed of three computational steps: (i) search for group-specific *k*-mers, (ii) consensus *k*-mer building, and (iii) coverage and specificity assessment.

Explorative Oligonucleotide Design) consists of three computational stages (Fig. 2.4). The user first provides two datasets that correspond to the target group and the non-target group. The first stage is the extraction of every *k*-mer from the target and the non-target groups using the Jellyfish program (Marcais and Kingsford, 2011). For large target groups containing more than 100 sequences, a noise reduction step is performed to remove untrustworthy *k*-mers that occur only once. Every *k*-mer found in both the target and the non-target groups is removed from the list of oligonucleotide candidates. The selected *k*-mers are then clustered together using CD-HIT (Li and Godzik, 2006) at an 88% identity threshold (i.e. allowing three mismatches for 25-mer probes). Only fully overlapping *k*-mers are clustered to gather *k*-mers from the same genomic location. For each cluster, a degenerate consensus is constructed that accounts for the sequence variability within the cluster. Among the combinations derived from each degenerate oligonucleotide, some correspond to sequences not previously included in the target group and therefore represent explorative probes.

Finally, the last stage of the KASpOD algorithm consists of assessing the coverage and specificity of each degenerate consensus *k*-mer. The coverage is evaluated against the target group using the PatMan program (Prüfer *et al.*, 2008), which allows the user to perform an exhaustive search with mismatches and indels to identify all occurrences of a high number of short sequences within a large database. The user defines the upper limit of tolerated mismatches. Specificity is assessed in the same way using the non-target group. KASpOD is provided as both a web service (http://g2im.u-clermont1.fr/kaspod/) and a stand-alone package. The software was used to design 25-mer probes for 1295 prokaryotic genera based on the recently published Greengenes taxonomy (McDonald *et al.*, 2012). The defined probe set allows each of the 252,183 high-quality and non-redundant 16S rRNA sequences to be covered by at least three different probes. Finally, 22,613 group-specific signatures were designed and are freely available on the KASpOD web site (http://g2im.u-clermont1.fr/kaspod/about.php). The alignment-free strategy allows computations to be completed in approximately two

weeks. Furthermore, this approach enables the definition of probes for large groups such as the *Corynebacterium* genus (20,093 sequences) where an alignment-based algorithm would have failed.

Functional gene arrays (FGAs)

Microbes mediate almost every conceivable biological process, and some researchers have estimated that individual environmental samples such as soil may contain between 10^3 and 10^7 different bacterial genomes (Curtis *et al.*, 2002, 2006; Gans *et al.*, 2005), each harbouring thousands of genes. In this context, high-density oligonucleotide FGAs provide the best high-throughput tools to access this tremendous diversity (He *et al.*, 2008). Currently, the most comprehensive FGA is the GeoChip (He *et al.*, 2007, 2010), which has evolved over several generations to be able to monitor most microbial functional processes, such as carbon, nitrogen, sulfur and phosphorus cycling, energy metabolism, antibiotic resistance, metal resistance, and organic contaminant degradation (He *et al.*, 2011, 2012a,b). Although most strategies are limited to the determination of probes that target specific gene sequences within a single genome dataset, few strategies offer the opportunity to design probes that permit broad coverage of multiple sequence variants for a given gene family (Tables 2.4–2.6) (Lemoine *et al.*, 2009; Dugat-Bony *et al.*, 2012b).

Probe design for FGAs using nucleic sequences

GeoChips are composed of 50-mer probes designed using a modified version of CommOligo (Li *et al.*, 2005). The experimental assessment of

Table 2.4 Comparison of probe design software features for functional gene arrays (FGAs): applications and availability [1]

Software	Reference	Application	Availability	URL
DEODAS (v 0.1.2)	Fredrickson *et al*. (2001)	FGA	Downloadable, GUI (L)	http://deodas.sourceforge.net/
Metabolic Design	Terrat *et al*. (2010)	FGA	Downloadable, GUI (W)	ftp://195.221.123.90/
ProDesign	Feng and Tillier (2007)	FGA	Web interface	http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html
ArrayOligoSelector (v 3.8.4)	Bozdech *et al*. (2003)	FGA, WGA-ORF	Downloadable, command-line (L)	http://arrayoligosel.sourceforge.net
CommOligo (v 2.0)	Li *et al*. (2005)	FGA, WGA-ORF	Downloadable, standalone GUI (W)	http://ieg.ou.edu/software.htm
HiSpOD	Dugat-Bony *et al*. (2011)	FGA, WGA-ORF	Web interface	http://g2im.u-clermont1.fr/hispod
MProbe (v 2.0)	Li and Ying (2006)	FGA, WGA-ORF	Downloadable, GUI (W)	http://www.biosun.org.cn/mprobe/
OligoArray (v 2.1)	Rouillard *et al*. (2003)	FGA, WGA-ORF	Downloadable, command-line (L)	http://berry.engin.umich.edu/oligoarray2_1/
OligoPicker (v 2.3.2)	Wang and Seed (2003)	FGA, WGA-ORF	Downloadable, command-line (L)	http://pga.mgh.harvard.edu/oligopicker/
OligoWiz (v 2.2.0)	Wernersson and Nielsen (2005)	FGA, WGA-ORF	Downloadable client program, GUI (L, W, M)	http://www.cbs.dtu.dk/services/OligoWiz2
PRIMEGENS (v 2.0)	Xu *et al*. (2002)	FGA, WGA-ORF	Web interface or command-line standalone (L, W)	http://primegens.org/
UPS 2.0	Chen *et al*. (2010)	FGA, WGA-ORF	Web interface	http://array.iis.sinica.edu.tw/ups/

FGA, functional gene array; WGA-ORF, open reading-frame oriented whole-genome array; GUI, graphical user interface; L, Linux; M, MacOS; W, Windows.

Table 2.5 Comparison of probe design software features for functional gene arrays (FGAs): main features

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
DEODAS (v 0.1.2)	Range chosen by the user	5′- to 3′-end	All probes reaching selection criteria	No	No	No	No	No	Yes
Metabolic Design	Fixed by the user	5′- to 3′-end	All probes reaching selection criteria	No	No	No	No	No	Yes
ProDesign	Range chosen the user (20–70)	5′- to 3′-end	Maximum number of probes chosen by the user	Yes	Yes	Yes	Yes	Yes	No
ArrayOligoSelector (v 3.8.4)	Fixed by the user	Probes ranking according to the 3′-end distance	Chosen by the user	Yes	Yes	Yes	No	No	No
CommOligo (v 2.0)	Fixed by the user (8–128)	Design starting from the 3′- or 5′-end	Chosen by the user	Yes	Yes	Yes	Yes	No	No
HiSpOD	Fixed by the user (18–120)	5′- to 3′-end	All probes reaching selection criteria	No	Yes	Yes	Yes	No	Yes
MProbe (v 2.0)	Range chosen by the user (20–100)	5′- to 3′-end	All probes reaching selection criteria	Yes	No	Yes	Yes	No	No
OligoArray (v 2.1)	Fixed by the user (15–75)	Distance to the 3′-end specified by the user (max. 1500)	Chosen by the user	Yes	Yes	Yes	Yes	No	No
OligoPicker (v 2.3.2)	Fixed by the user (20–100)	Design chosen for the 5′- or the 3′-end	Chosen by the user (up to 5)	Yes	Yes	No	Yes	No	No
OligoWiz (v 2.2.0)	Fixed by the user	Localization score based on centre, 5′ or 3′ distance	All probes reaching selection criteria	Yes	No	No	Yes	No	No
PRIMEGENS (v 2.0)	Fixed by the user	No localization specified	Chosen by the user	Yes	No	Yes	Yes	No	No
UPS 2.0	Fixed by the user (20–120)	5′- to 3′-end	Chosen by the user (up to 10)	Yes	Yes	Yes	Yes	Yes	Yes

Table 2.6 Comparison of probe design software features for functional gene arrays (FGAs): flexibility

Software	Organism	Cross-hybridization assessment	Database for specificity test	Input files
DEODAS (v 0.1.2)	No limitation	EMBOSS	GenBank	A FASTA file containing targeted sequences. Protein sequences.
Metabolic Design	No limitation	BLAST and Kane's specifications	EnvExBase (10GB) Complete CDS database	No input files are required
ProDesign	No limitation	Spaced seed hashing and Kane's specifications	Input sequence dataset	A FASTA file containing targeted sequences and optionally a cluster file. Nucleotide sequences
ArrayOligoSelector (v 3.8.4)	No limitation	BLAST and thermodynamic calculations	External FASTA file (typically single organism genome)	Two FASTA files (targeted sequences and the complete genome). Nucleotide sequences
CommOligo (v 2.0)	No limitation	Global alignment and thermodynamic calculations	Input sequence dataset	A FASTA file containing targeted sequences. Nucleotide sequences.
HiSpOD	No limitation	BLAST and Kane's specifications	EnvExBase (10GB) Complete CDS database	A FASTA file containing targeted sequences. Consensus or non-degenerate nucleotide sequences
MProbe (v 2.0)	No limitation	BLAST and Kane's specifications	Input sequence dataset	A GenBank, EMBL or FASTA file containing targeted sequences. Nucleotide sequences
OligoArray (v 2.1)	No limitation	BLAST and thermodynamic calculations	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences
OligoPicker (v 2.3.2)	No limitation	BLAST	Input sequence dataset or external FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences
OligoWiz (v 2.2.0)	All organisms found on the server	BLAST, Kane's specifications and thermodynamic calculations	Single organism genome (among a list of organisms found on the server)	A FASTA file containing targeted sequences or a tab-delimited file containing both sequences and annotations. Nucleotide sequences
PRIMEGENS (v 2.0)	No limitation	BLAST, Kane's specifications and multiple sequence alignment	External FASTA file or a complete genome sequence among a list of organisms found on the server	A FASTA file containing targeted sequences. Nucleotide sequences
UPS 2.0	No limitation	BLAST and thermodynamic calculations	Input sequence dataset, complete genome sequence among a list of organisms found on the server, NCBI Nucleotide database or external FASTA file	A FASTA file containing targeted sequences and optionally a fasta file with non-target sequences. Nucleotide sequences

optimal probe design criteria (He *et al.*, 2005) permitted CommOligo to be implemented to combine three different parameters for sensitivity and specificity evaluation: sequence identity, free energy and continuous stretch. For each sequence, the first stage consists of masking oligonucleotides according to different filters including distance to the 3′ untranslated region (UTR), GC content, complexity, degeneracy and specificity (i.e. significant matching of oligonucleotides with non-targets). Continuous matches of a user-specified length with non-targets are assessed using an algorithm similar to that of OligoPicker (Wang and Seed, 2003) by storing all possible 10-mers within the sequences in a hash table data structure. Thus, the hash key is a 10-mer sequence, and the hash value corresponds to the relative sequence indices and positions where this particular 10-mer is found. Strictly identical 10-mers shared between probes and non-targets are not retained. This data structure is also used to assess the self-annealing of each unmasked oligonucleotide by searching for continuous matches of a user-defined length within the tested oligonucleotide itself. Probes that show self-annealing are filtered out. The remaining probes are tested for specificity against non-targets using both a global alignment algorithm (Myers, 1999) and a binding free energy calculation rather than the classically used Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990). Sequence identity is therefore inferred from the percentage of matches in a global gapped alignment, and oligonucleotides with high identity (i.e. higher than a cut-off value defined by the user) to non-targets are filtered out. Oligonucleotides with medium identity but low free energy are also removed. Then, the program computes the best interval of melting temperatures that covers most targets and probes and removes all candidate oligonucleotides outside of this range. Finally, a sequence may have more probes than needed, in which case CommOligo is able to select oligonucleotides using a multi-criterion optimization algorithm where cross-hybridization, positions and identity between probes are taken into account. Gene-specific probes can be selected using CommOligo with the following parameters: < 90% sequence identity, <20-base continuous stretch, and >−35 kcal/mol free energy with non-targets (Liebich *et al.*, 2006; He *et al.*, 2011). Additionally, a group-specific probe design can be performed by adding these supplemental criteria: > 96% sequence identity, >35-base continuous stretch, and <−60 kcal/mol free energy within the targeted group (He *et al.*, 2005, 2011). CommOligo performs complex and time-consuming calculations. However, the version available for download is not well suited to conducting high-throughput analyses, and with the increasing availability of sequences corresponding to protein-coding genes (complete genome sequencing and environmental studies from specific functional markers), new software has been developed in the last decade that takes this wide diversity into account.

Hierarchical Probe Design (HPD) software (Chung *et al.*, 2005) was the first program dedicated to FGAs that was based on the concept of cluster-specific probes. The first step of the algorithm consists of the multiple sequence alignment of input sequences using ClustalW (Thompson *et al.*, 1994). A hierarchical clustering is then performed using either a neighbour-joining (Saitou and Nei, 1987) or a UPGMA (Sokal and Michener, 1958) method. All candidate probes are subsequently generated, and cluster-specific probes are selected using a bottom-up approach (Fig. 2.5). The specificity of candidate oligonucleotides is checked against clusters that are one level higher. If a probe of one sibling cluster harbours sufficient specificity to discriminate among these clusters, it remains in the sibling cluster. If not, the candidate is transferred to the upper cluster and therefore represents a group-specific probe. This recursive process is repeated as long as the root cluster has not been reached. The optimal probe set is then determined according to probe quality criteria including cluster coverage, specificity, GC content and hairpin energy. Although this tool is not explorative, it automatically produces probes against all nodes of the clustering tree, thereby providing extensive coverage of known variants from a conserved functional gene. However, at this time, the software no longer appears to be available. ProDesign (Feng and Tillier, 2007) uses similar clustering methods with the aim of detecting all members of a gene family in environmental samples. However, in contrast to HPD, this

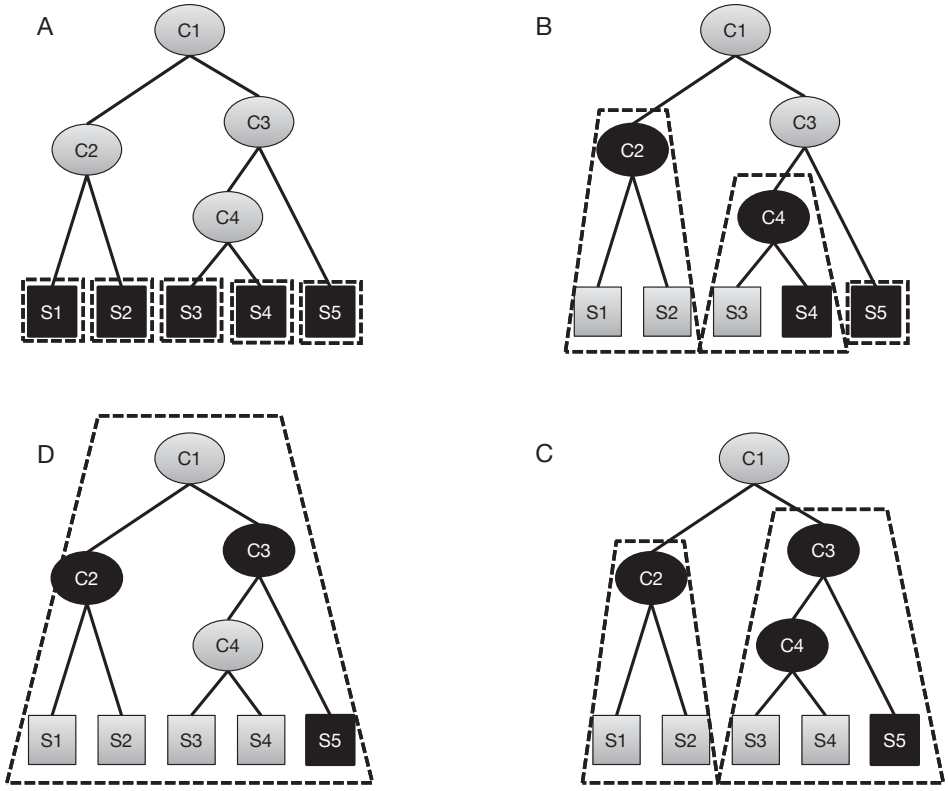


Figure 2.5 HPD probe candidate selection process. Sequences are hierarchically clustered (A), and a bottom-up approach (B and C) is performed to search for putative group-specific probes that can target the whole dataset (D). Black circles or squares within dotted regions indicate that probe candidates exist for that cluster or sequence. Grey circles or squares within the dotted regions indicate that no probe candidate exists for that cluster or sequence. The circles outside of the dotted region indicate the clusters that have not yet been explored.

software uses sophisticated spaced seed hashing rather than a suffix tree algorithm to benefit from permitted mismatches between a probe and its targets, and it ensures the re-clustering of groups for which no probe was found, which results in a significant improvement in sequence coverage. Although both of these strategies allow the coverage of a wider range of sequence variants, they only permit the survey of known sequences and therefore cannot be used to evaluate the unknown microbial genes present in complex environments. The main drawbacks of these strategies are thus their inability to generate explorative probes and the absence of an evaluation of specificity (i.e. searching for potential cross-hybridizations) against large databases that are representative of microbial diversity.

To overcome these limitations, the HiSpOD (High Specific Oligo Design) program was developed (Dugat-Bony *et al.*, 2011) in the context of microbial ecology. HiSpOD includes the classical parameters for the design of effective probes, including probe length, melting temperature, GC content and complexity, and adds supplemental properties that were not considered by previous programs. HiSpOD allows the design of degenerate probes for gene families after multiple alignments of nucleic sequences belonging to the same gene family and can produce consensus sequences. All combinations deduced from the degenerate probes are then divided into two groups (Fig. 2.6A). The first group corresponds to specific probes for sequences available in databases, and the second group corresponds to explorative probes that represent putative new signatures that do not correspond to any previously described microorganisms. A probe set representing the most likely gene sequence variants and a probe set representing new combinations that have not yet been deposited in databases are created based on multiple mutation events that have already been identified. Sequence-specific probes

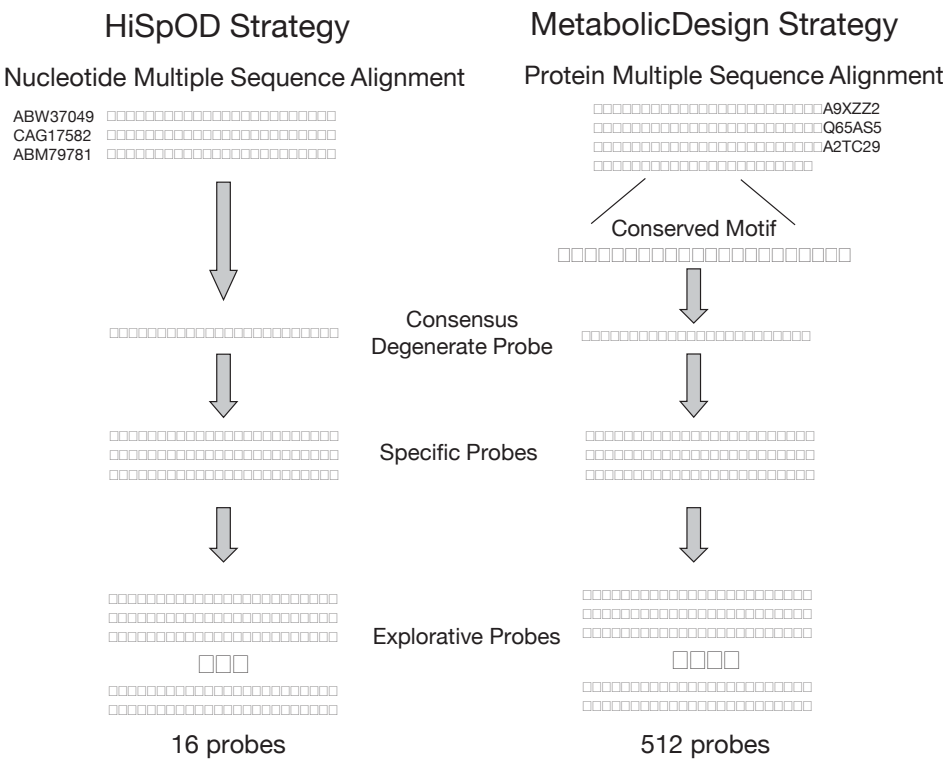


Figure 2.6 Comparison of the explorative probe design strategies implemented in HiSpOD and Metabolic Design software. The example shows the probe design for the *bphA1c* gene encoding the salicylate 1-hydroxylase alpha subunit involved in PAH degradation from three distinct *Sphingomonas* or *Sphingobium* species using both strategies.

can also be designed through HiSpOD by using non-degenerate classical nucleic acid sequences. To limit cross-hybridization, the specificity of all selected probes is checked against a large formatted database dedicated to microbial communities, i.e. the EnvExBase (Environmental Expressed sequences dataBase), which is composed of all coding DNA sequences (CDSs) from the prokaryotic (PRO), fungal (FUN) and environmental (ENV) taxonomic divisions of the EMBL databank. Specificity tests are performed using BLAST (Altschul *et al.*, 1990), and cross-hybridization results are clustered using a single-linkage method implemented in BLASTCLUST (Altschul *et al.*, 1990).

Probe design for FGAs using protein sequences

In contrast to the strategies outlined above, several new strategies have been proposed to initiate probe design from conserved peptidic regions rather than from nucleic acid sequences to survey all potential nucleic acid variants.

The first strategy based on this principle was described by Bontemps *et al.* (Bontemps *et al.*, 2005) and called CODEHMOP (Consensus Degenerate Hybrid Motif Oligonucleotide Probe). This strategy is derived from an adaptation of the CODEHOP (Consensus Degenerate Hybrid Oligonucleotide Primer) PCR primer design strategy, which was originally developed to identify distantly related genes encoding proteins that belong to known families (Rose *et al.*, 1998, 2003; Boyce *et al.*, 2009). The CODEHMOP strategy aims to identify conserved amino acid motifs from multiple alignments of protein sequences. Then, the most highly conserved region (5–7 amino acids) of each protein motif is back-translated to generate all possible nucleic combinations (15–21 nucleotides) coding for this peptide. These sequences are extended by 5′ and 3′ fixed ends (12–15 nucleotides each) that are derived from the most frequent nucleotide at each position flanking the conserved region in the nucleotide sequence alignment. The final probes are called ‘hybrids’, as they comprise a variable

central core with some nucleic combinations that do not correspond to any sequences yet described (to target greater diversity) combined with two fixed end sequences (available in databases) that are added to increase the probe length. This approach was used to design a prototype DNA array that included all described and undescribed *nodC* (nodulation gene) sequences in bacteria and that was applied to legume nodule samples (Bontemps *et al.*, 2005). This strategy enabled the detection of new *nodC* sequences that exhibited less than 74% identity with known sequences. The application of the CODEHMOP strategy is, however, limited by its lack of implementation in a fully automated program and its lack of probe specificity test. Nevertheless, this approach appears to be the most comprehensive way to encompass the diversity of gene sequence variants potentially found for enzymes mediating a given function.

Terrat *et al.* (2010) developed a new software program called Metabolic Design that ensures the *in silico* reconstruction of metabolic pathways, the identification of conserved motifs from multiple protein alignments, and the generation of efficient explorative probes through a simple convenient graphical interface. In this case, before the probe design stage, the user reconstructs the chosen metabolic pathway *in silico* with all substrates and products from each metabolic step. One reference enzyme for each of these steps is selected, and its protein sequence is extracted from a curated database (by default, Swiss-Prot), which is then used to retrieve all homologous proteins from complete databases (Swiss-Prot and TrEMBL). After the most pertinent homologous sequences are selected, they are aligned to begin the probe design stage. The amino acids are back-translated for each identified molecular site, with all redundancy of the genetic code taken into account, to produce a degenerate nucleic consensus sequence. All degenerate probes that meet the criteria defined by the user (probe length and maximal degeneracy) are retained. All of the possible specific combinations for each degenerate probe are subsequently checked for potential cross-hybridization against a representative database (e.g. EnvExBase as in the HiSpOD program). Finally, an output file listing all of the degenerate probes selected by the user permits the deduction of all possible combinations and organizes them into specific probes and exploratory probes (Fig. 2.6B).

Whole-genome arrays (WGAs)

Many organisms that are closely related based on SSU rRNA gene sequences can exhibit remarkably different phenotypic characteristics that result from great differences in their genomes, which in turn arise from processes such as lateral gene exchange (Gentry *et al.*, 2006). Whole-genome arrays that use whole-genome sequence information of one or several closely related microorganisms provide a way of understanding such phenotypic differences (Zhou, 2003). WGAs are divided into two main groups: whole-genome ORF arrays, which contain oligonucleotide probes for all of the open reading frames (ORFs) in a genome, and tiling arrays, which represent a complete non-repetitive tile path over the genome, irrespective of any genes that may be annotated in a particular region (Bertone *et al.*, 2006).

Whole-genome ORF arrays

Probe design considerations for whole-genome ORF arrays are similar to those for functional gene arrays (FGAs). However, some algorithms are specifically dedicated to the design of oligonucleotide probes for whole-genome ORF arrays. One of the most-cited software programs for designing such microarrays is PICKY (Chou *et al.*, 2004b). PICKY was initially developed for oligonucleotide microarray design for large eukaryotic genomes and thus boasts major speed improvements when compared with other whole-genome ORF array probe design programs. Several probe design criteria are considered by PICKY to compute the optimal probe set, such as the complexity (i.e. no single base should constitute more than 50% of a probe and no stretch of the same base should exceed 25% of the length of a probe), thermodynamic criteria (i.e. a GC content between 30% and 70% and no secondary structures), and cross-hybridization [i.e. Kane's criteria (Kane *et al.*, 2000)]. For the latter criterion, PICKY can handle multiple target and non-target gene sets; thus, oligonucleotide probes are defined for the target set and to prevent hybridization with the

non-target set. Most of the previously mentioned criteria are user-adjustable through a user-friendly graphical interface. Finally, to ensure the uniformity of the probe set, PICKY is able to adjust the probe length within a user-defined range.

PICKY relies on the construction of a generalized suffix array where both strands are represented. The suffix array is built using a modified Burkhardt-Kärkkaïnen algorithm (Burkhardt and Kärkkäinen, 2003) that allows quick and efficient construction. Using this suffix array, PICKY can first exclude low-complexity and repetitive genomic regions as well as self-similar and self-complementary regions for probe design. Such screening allows the detection of putative secondary structures without using dedicated external software. PICKY then avoids other unnecessary computations by removing regions that fail to comply with Kane's criteria, as these regions may cross-hybridize with non-target sequences. For all remaining regions, PICKY computes a score to indicate the likelihood of cross-hybridization and then prioritizes regions for oligonucleotide selection. Once all probe candidates have been computed, the melting temperatures for all possible probe/target and probe/non-target pairs are estimated according to Kane's second condition, which states that any sequence similarity over 75% identity (or a user-defined value) can potentially involve cross-hybridization. The melting temperature of each candidate probe is assessed against its target, and all non-targets are gathered using the suffix array. As probe/non-target pairs may have mismatches, melting temperatures are not precise but are sufficient to predict whether such duplexes will potentially be present.

The calculated melting temperatures of candidate probes with all of its non-targets are then used to prioritize probes for the final processing step. This last step consists of multi-objective optimization to compute the probe set best able to detect each gene, i.e. by avoiding cross-hybridization and sharing a uniform melting temperature range. Thus, PICKY is a fast and efficient probe design software program for whole-genome ORF arrays that addresses numerous design criteria to compute the optimal probe set. No additional external software is required to run PICKY, and it is easy to use through a graphical user interface that is available for all major computing platforms (Mac, Windows and Linux).

Several existing probe design software programs (Tables 2.7–2.9) are free to use and still available for designing whole-genome ORF arrays with various strategies such as Mprime (Rouchka *et al.*, 2005), OliD (Talla *et al.*, 2003), PROBESEL (Kaderali and Schliep, 2002) or ProbeSelect (Li and Stormo, 2001).

Targeting all genes through the use of whole-genome ORF arrays may not be sufficient for some applications, such as the identification of transcription in the antisense strand or regulatory pathway discovery (Bertone *et al.*, 2005). Consequently, PICKY proposes PERL scripts for tiling array purposes.

Table 2.7 Comparison of probe design software features for whole-genome ORF arrays (WGAs): applications and availability

Software	Reference	Application	Availability	URL
Mprime	Rouchka *et al*. (2005)	WGA-ORF	Web interface	http://kbrin.a-bldg.louisville.edu/Tools/OligoDesign/MPrime.html
OliD	Talla *et al*. (2003)	WGA-ORF	Downloadable, command line (L)	Upon request
PICKY (v 2.2)	Chou *et al*. (2004b)	WGA-ORF	Downloadable, standalone GUI (L, W, M)	http://www.complex.iastate.edu./download/Picky/index.html
ProbeSelect	Li and Stormo (2001)	WGA-ORF	Available upon request, command line (L)	http://stormo.wustl.edu/src/probeselect-src.tar

WGA-ORF, open reading-frame oriented whole-genome array; GUI, graphical user interface; L, Linux; M, MacOS; W, Windows.

Table 2.8 Comparison of probe design software features for whole-genome ORF arrays (WGAs): main features

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
Mprime	Fixed by the user	Probes weighted towards 3′-end	Chosen by the user	Yes	No	Yes	Yes	No	No
OliD	Fixed by the user	Preference given to the 3′-end proximity	Chosen by the user	Yes	Yes	Yes	No	No	No
PICKY (v 2.2)	Range chosen by the user (50–90)	No localization specified	Chosen by the user (up to 5)	Yes	Yes	Yes	Yes	No	No
ProbeSelect	Fixed by the user	No localization specified	Chosen by the user	Yes	Yes	No	Yes	Yes	No

Table 2.9 Comparison of probe design software features for whole-genome ORF arrays (WGAs): flexibility

Software	Organism	Cross-hybridization assessment	Database for specificity test	Input files
Mprime	Rat, mouse, human, drosophila and zebrafish	BLAST (wuBLAST)	RefSeq database for the organism	Gene name, GenBank accession number, keyword, or FASTA files. Nucleotide sequences.
OliD	No limitation	BLAST	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences.
PICKY (v 2.2)	No limitation	Suffix array approach, Kane's specifications and thermodynamic calculations	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
ProbeSelect	No limitation	Suffix array approach and thermodynamic calculations	Single organism genome	A FASTA file containing targeted sequences. Nucleotide sequences.

Tiling arrays

In contrast to whole-genome ORF arrays, tiling arrays aim to determine probes over the whole genome irrespective of any genes that may be annotated in the genome. The design of oligonucleotide tiling arrays is different from the selection of oligonucleotides for gene-based arrays, and additional factors should be considered, such as tiling resolution and the handling of non-unique subsequences.

A naïve strategy for selecting oligonucleotides for a whole-genome tiling array is to generate a tile path from the beginning of a chromosome to its end and cover the entire sequence with 25-mer probes tiled end-to-end (Yamada *et al.*, 2003). Many of the probes chosen using this approach may, however, be subject to hybridization problems and potentially result in the misinterpretation of results. Some probes may be redundant, some may be thermodynamically unable to hybridize and some may be non-specific and thus undergo cross-hybridization (Mockler *et al.*, 2005). Consequently, in most cases, such a naive strategy is not the optimal approach to tiling, and the inclusion

of criteria such as tiling resolution and repetitive region masking is essential to ensure the best probe design for tiling arrays.

Tiling density (Fig. 2.7) is an important factor in a tiling array design because it determines how the genome should be subdivided and how densely oligonucleotide probes are placed. Probes can be contiguous (i.e. tiled end-to-end) or discontiguous, including gaps with a predetermined size range between probes for single-copy tiling. For some applications, it may be necessary to have a better resolution involving multiple feature tiling. The whole genome is covered with multiple oligonucleotide probes such that the starting position of each probe is shifted by one or several nucleotides to overlap the previous oligonucleotide's coordinates (Bertone *et al.*, 2006). Although such a strategy allows a fine-resolution analysis, the number of probes determined will eventually dramatically increase. However, advances in high-resolution microarray technology have enabled the inclusion of up to 4.2 million probes on an array.

Because genome sequences are not random, many redundant subsequences are scattered all along the genome. Therefore, once the tiling strategy has been determined, a common first step is to perform a genomic repeat masking prior to probe selection. To obtain an easy and relevant interpretation of a tiling array experiment, it is necessary to avoid the generation of multiple oligonucleotide probes sharing the same sequence. Because such sequences generally correspond to known repetitive sequences, algorithms such as RepeatMasker (http://www.repeatmasker.org/) are widely used to easily address this problem. RepeatMasker is capable of identifying genomic repeats in a variety of genomes using a database of well-characterized families of repetitive elements (Jurka *et al.*, 2005). It is also preferable to remove low-complexity DNA regions (i.e. stretches of the same nucleotide or regions with extremely high A/T or G/C content). RepeatMasker allows the filtering of some low-complexity sequences by default, but it could be necessary in some cases to combine RepeatMasker with dedicated software that calculates entropy scores, such as NSEG (Wootton and Federhen, 1993) or DUST (Hancock and Armstrong, 1994), for a more intensive filtering, especially of specific repetitive sequences of the studied genome.

The next step consists of filtering out oligonucleotide probes based on thermodynamic considerations (SantaLucia, 1998). Low-binding affinity probes are useless in a tiling array experiment, as are high affinity non-specific probes, which would be uninformative because of the saturated cross-hybridization. Therefore, probe

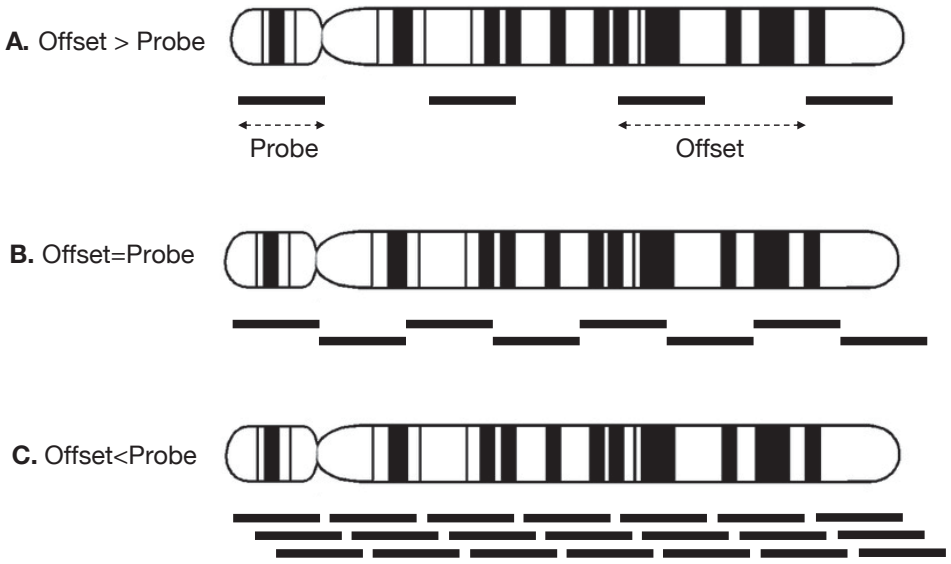


Figure 2.7 Tiling density. The offset between probes is the distance between the start of one probe and the start of the next. Three different tiling densities are shown: (A) illustrates gapped tiling, (B) end-to-end tiling and (C) overlapping tiling.

affinity modelling and the determination of probe specificity on a whole-genome level can be used to screen candidate oligonucleotides and eliminate those likely to be problematic from the microarray design (Mockler *et al.*, 2005).

For all of these reasons, it is necessary to adapt the tiling strategy and placement of oligonucleotide probes along the genome to obtain the optimal probe set. To design the oligonucleotides in each sequence window, the probe design software has to address position and hybridization quality (Lemoine *et al.*, 2009). Tables 2.10–2.12 summarize the available and free-to-use probe design software dedicated to tiling arrays: chipD (Dufour *et al.*, 2010), Teolenn (Jourdren *et al.*, 2010), MOPeD (Patel *et al.*, 2010), PanArray (Phillippy *et al.*, 2009), Tileomatic (Schliep and Krause, 2008), ArrayDesign (Gräf *et al.*, 2007) and MAMMOT (Ryder *et al.*, 2006).

Among these algorithms, Teolenn appears to be the most universal and flexible software to address the tiling array design problem and remains easy to use despite its command-line utilization. Teolenn relies on a four-step workflow where each step is customizable. Thus, users are allowed to activate or deactivate each function according to their needs or available computational resources. Teolenn accepts both masked (e.g. using RepeatMasker) and unmasked genome sequences in FASTA format as input.

The first step consists of generating all possible non-redundant oligonucleotide probes along the whole genome. Probe length can be fixed or may vary within a user-defined length range. In the second step, for all created probes, Teolenn assesses the oligonucleotide quality based on several criteria including melting temperature, GC content, complexity and uniqueness. Melting temperatures are computed using a nearest-neighbour method (SantaLucia, 1998), complexity is measured by counting the masked bases and uniqueness within the genome is evaluated according to Gräf *et al.* (2007). The third step of Teolenn is probe filtering, which strongly depends on the tiling array application. For example, if the user needs a transcriptome array, Teolenn is able to filter out all probes that are not located within an ORF as well as small RNAs based on genome annotations. However, if a homogeneous tiling path along the genome is desired, Teolenn can keep all of the possible probes without stringent quality filters. Eventually, the best probe in each genomic window that maximizes a position score, where the most central probe has the best score, and a previously assessed quality score can be selected. Depending on the user's needs, all score calculations can be weighted. The designed probes can be output in plain text, FASTA format or GFF. The GFF format allows the visualization of the results in a genome browser such as gBrowse (Stein *et al.*, 2002). Teolenn therefore provides a complete and flexible software program dedicated to tiling arrays.

Table 2.10 Comparison of probe design software features for tiling arrays: applications and availability

Software	Reference	Application	Availability	URL
ChipD	Dufour *et al*. (2010)	WGA-tiling	Web interface	http://chipd.uwbacter.org/
MAMMOT (v 1.21)	Ryder *et al*. (2006)	WGA-tiling	Downloadable, local server	http://www.mammot.org.uk/
MOPeD	Patel *et al*. (2010)	WGA-tiling	Web interface	http://moped.genetics.emory.edu/newdesign.html
OligoTiler	Bertone *et al*. (2006)	WGA-tiling	Web interface	http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi
PanArray (v 1.0)	Phillippy *et al*. (2009)	WGA-tiling	Available upon request, command line (L)	Upon request
Teolenn (v 2.0.1)	Jourdren *et al*. (2010)	WGA-tiling	Downloadable, command line (L)	http://transcriptome.ens.fr/teolenn/

WGA-tiling: tiling whole-genome array. L: Linux.

Table 2.11 Comparison of probe design software features for tiling arrays: main features

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
ChipD	Range chosen by the user (greater than 15)	Read input sequences from 5′-end to 3′-end	Maximum number of probes chosen by the user	No	Yes	No	Yes	Yes	No
MAMMOT (v 1.21)	Fixed by the user	Start and end locations are chosen by the user	All probes reaching selection criteria	No	Yes	Yes	Yes	No	No
MOPeD	Range chosen by the user (55–65)	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	Yes	No	No	Yes	No	No
OligoTiler	Fixed by the user	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	Yes	No	No	No	No
PanArray (v 1.0)	Fixed by the user	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	No	No	No	No	No
Teolenn (v 2.0.1)	Fixed by the user or range chosen by the user	Read input sequences from 5′-end to 3′-end	All probes reaching selection criteria	No	Yes	Yes	Yes	No	No

Table 2.12 Comparison of probe design software features for tiling arrays: flexibility

Software	Organism	Cross-hybridization assessment	Database for specificity test	Input files
ChipD	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
MAMMOT (v 1.21)	No limitation	No	No	A FASTA file with the complete genome sequence. Nucleotide sequences.
MOPeD	Human, mouse, rhesus monkey	No	No	No input files are required.
OligoTiler	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
PanArray (v 1.0)	No limitation	No	No	A FASTA file containing targeted sequences (typically multiple genome sequences). Nucleotide sequences.
Teolenn (v 2.0.1)	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.

Other applications

Transcriptome arrays

Gene-oriented whole-genome arrays and tiling arrays can both be used to measure the expression of thousands of genes of an organism, thus providing a snapshot of the transcriptome in different states in tissues and cells (Nakaya *et al.*, 2007). Compared with tiling arrays, gene-based WGAs appear to be relatively simple to handle, as they use fewer probes for each gene. In contrast, tiling arrays are capable of providing information regarding alternative splicing or realizing the transcriptome annotation of a newly sequenced organism.

The choice between a gene-oriented WGA and a tiling platform is highly correlated with the biological question. There are also many commercial gene expression microarrays available that should be considered before undertaking a custom design.

Typing microarrays

Identifying microbial communities using microarrays is a common task that is generally performed using probe design software dedicated to POAs. However, for some applications such as discrimination among several strains of the same bacterial species or the classification of plasmids, it is necessary to develop dedicated tools. Such tools (Meng *et al.*, 2008; Vijaya Satya *et al.*, 2008) aim to design an optimal set of oligonucleotide fingerprints that is able to distinguish among similar targets. For particular applications including strain detection microarrays, a custom design is often inevitable. Two different probe design approaches are used to build these microarrays: a traditional gene-oriented approach and a strategy originally developed for sequencing by hybridization: resequencing microarrays (Leski *et al.*, 2012). Oligonucleotide probe design for resequencing microarrays is substantially different from previously discussed strategies. Such strategies were developed to characterize bacterial pathogens by sequencing a significant portion of their genome or, for viral pathogens, the whole genome. Multiple versions of each oligonucleotide are identified, i.e. four probes in both the sense and antisense directions, for a total of eight probes per base. The four probes differ by only one central nucleotide to represent the four possible base combinations (i.e. A, T, C or G). Based on hybridization results, base-calling algorithms are able to obtain a reliable sequence that can be compared with public sequence databases (Fig. 2.8).

Fingerprints could also be generated by random probes (Belosludtsev *et al.*, 2004). In this example, the authors created 9015 12-mers with homogeneous GC contents where each probe differed by at least four bases and 5268 13-mers that differed by at least five bases. After the hybridization step, the authors demonstrated that their strategy produced reproducible patterns of hybridization that distinguished among species. The use of such a probe design eliminates the need for updating the array design when new bacterial genomes become available. The primary drawback of this strategy is the need to experimentally obtain the hybridization pattern for organisms without known sequences.

Discussion/challenges and future trends

The success of a microarray experiment strongly depends on the determination of the best probe set while taking the biological question into account. Despite the development of numerous probe design strategies, some parameters require particular attention, as they have significant impact on probe specificity, sensitivity and quantitative capability (Zhou, 2003; Wagner *et al.*, 2007).

Explorative probe design

For environmental DNA microarrays including phylogenetic oligonucleotide arrays (POAs) and functional gene arrays (FGAs), explorative probe design strategies offer the opportunity to survey both known and unknown microorganisms (Dugat-Bony *et al.*, 2012b). Explorative probes use the sequence variability within the targeted sequences to define new combinations that have not yet been deposited in public databases but are potentially present in the environment. One future development in microarrays will be to incorporate the original concept into probe design software, especially by offering the ability to design group-specific degenerate probes.

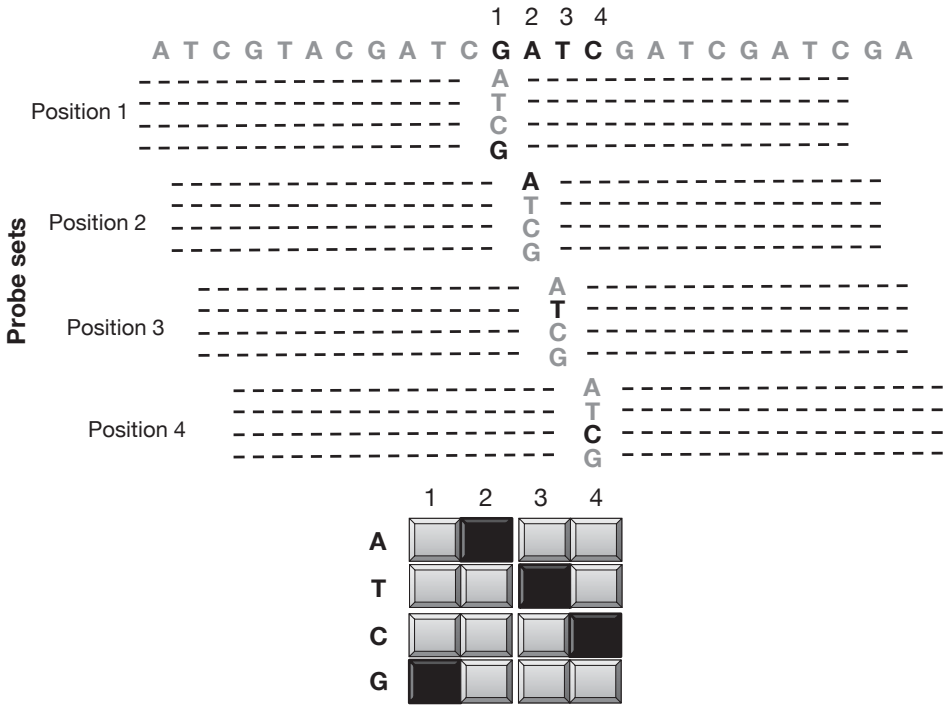


Figure 2.8 Resequencing microarrays. An example of a probe set construction targeting four contiguous nucleotides (1, 2, 3 and 4) of a prototype sequence using Affymetrix technology. Overlapping sets of probes shifted by one nucleotide covering the whole prototype sequence are generated. Each probe set contains one matching (black) and three mismatched (grey) probes that differ only by the central nucleotide (four additional probes could be designed for the other strand). Black lines represent sequences with the prototype sequence above. The bottom figure shows a scan of a segment of the microarray to which the labelled targets were hybridized. Black squares correspond to a positive signal. The base calls are therefore made according to the hybridization pattern.

Probe length

Another probe design criterion that impacts both sensitivity and specificity is probe length. Short oligonucleotide probes are more specific but less sensitive than long probes (Guschin *et al.*, 1997). The building of phylogenetic oligonucleotide microarrays requires the determination of short fingerprints that are able to discriminate among microbial taxa. Existing POAs, therefore, use short probes (i.e. 24- to 25-mers) (Brodie *et al.*, 2006; Rajilić-Stojanović *et al.*, 2009; Hazen *et al.*, 2010; Handley *et al.*, 2012; Paliy and Agans, 2012), whereas FGAs or whole-genome ORF arrays may be built with either short (i.e. 15–30-mers) (Bodrossy *et al.*, 2003; Stralis-Pavese *et al.*, 2004) or long oligonucleotides (i.e. 40- to 70-mers) (Kane *et al.*, 2000; Relógio *et al.*, 2002; He *et al.*, 2007, 2010; Dugat-Bony *et al.*, 2012a). The primary limitation of microarrays based on short oligonucleotide probes is the need to use, in most cases, PCR-amplified targets to ensure enrichment, which introduces an inherent PCR bias (Suzuki and Giovannoni, 1996; Vora *et al.*, 2004).

A promising alternative approach to the design of oligonucleotide probes is the use of the GoArrays strategy (Rimour *et al.*, 2005) (http://g2im.u-clermont1.fr/serimour/goarrays.html). Such a strategy enables the production of oligonucleotide probes that are as specific as short probes and as sensitive as long probes, and consists of the concatenation of two short subsequences that are complementary to disjoined regions of the target, with an insertion of a short random linker (i.e. 3–6-mers). This strategy has been shown to improve microarray efficiency for a wide range of applications (Rimour *et al.*, 2005; Zhou *et al.*, 2007; Pariset *et al.*, 2009; Kang *et al.*, 2010).

Databases

Most currently available probe design software programs only perform specificity tests against a

reduced set of sequences, such as whole-genome data or specific sets of genes (Lemoine *et al.*, 2009). Environmental DNA microarrays, however, require dedicated datasets that are as representative as possible of all of the non-target sequences potentially present in the samples. Because only a small portion of the total natural microbial diversity has been documented, it is a major challenge to design suitable probes that are specific to unique markers and do not cross-hybridize with putative and currently unknown similar sequences (Chandler and Jarrell, 2005). There is a trade-off between using the largest databases and thus minimizing putative cross-hybridizations and using small, dedicated databases that are less time-consuming for specificity tests. The major public databases including GenBank (Benson *et al.*, 2012), the European Nucleotide Archive (ENA) (Leinonen *et al.*, 2011) and the DNA Data Bank of Japan (DDBJ) (Kaminuma *et al.*, 2011) are the most complete nucleic sequence databases with which to perform specificity tests. However, the use of such databases could drastically increase the run times of probe design software. In addition, for environmental DNA microarrays, entire public databases are not really appropriate because they contain some subsets that are not typically considered in microbial ecology, such as *Metazoa*. Furthermore, numerous erroneous annotations could negatively impact the quality of the probe design.

Within POAs, each probe must be specific with respect to all small subunit (SSU) rRNA sequences that may be present in the sample during hybridization. Curated and dedicated secondary databases that gather all of the SSU rRNA sequences described in public databases have already been constructed [e.g. Ribosomal Database Project (Cole *et al.*, 2009), Greengenes (DeSantis *et al.*, 2006b) and SILVA (Pruesse *et al.*, 2007)]. The differences among these databases arise from the construction and update workflows, which lead to distinct sizes: SILVA (Release 111) contains 3,194,778 16S rRNA sequences, RDP (Release 10) contains 2,578,902 16S rRNA sequences and Greengenes (10/2/2011) contains 1,049,116 16S rRNA sequences. Because they are smaller and contain no unnecessary data, these databases are well-adapted to the construction of prokaryotic POAs. PhylArray software (Militon *et al.*, 2007) was, however, developed before these databases were publicly available, and therefore uses its own highly curated (full-length and quality-filtered) and automatically updated prokaryotic SSU rRNA database (66,076 sequences for the last release).

For environmental FGAs, the database used for specificity tests must include all known CDSs that may be encountered in natural environments. To the best of our knowledge, EnvExBase (used in the HiSpOD and Metabolic Design programs) is the first CDS database dedicated to microbial ecology (Terrat *et al.*, 2010; Dugat-Bony *et al.*, 2011). For its construction, all annotated transcript sequences and their associated 5′ and 3′ untranslated regions (UTR) in all classes of the EMBL prokaryotic (PRO), fungal (FUN) and environmental (ENV) taxonomic divisions were extracted and curated to remove low-quality sequences. EnvExBase thus represents a 13,697,580 sequence database.

Updates and performance

The constant increase in available sequences (Cochrane *et al.*, 2009) requires that databases for specificity tests must be regularly updated. As a result, probe datasets must be re-computed as frequently as possible to include all deposited data. However, as mentioned above, assessing probe specificity against large databases can be a time-consuming task in the probe design step. A complete 16S rRNA sequence database is approximately two million sequences, and a complete CDS database such as EnvExBase represents 14 million sequences. To overcome this limitation, an interesting strategy would be to create databases specific to each ecological compartment. Usually, specificity tests are not performed against a suitable subset of sequences, primarily because of the lack of databases for microbial ecology. Depending on the environment studied, it would be more relevant to perform these tests against reduced databanks dedicated to specific ecosystems (e.g. soil, marine, freshwater, and gut). However, for 'universal' tool development relevant to various environments, the most complete database must be considered.

The rapid growth of datasets, particularly environmental datasets, has led to an important

increase in computational requirements coupled with a fundamental change in the way that algorithms are conceived and designed [e.g. mpiBLAST (Darling *et al.*, 2003) or GPU-BLAST (Vouzis and Sahinidis, 2011)]. Efforts to limit computation time are based on exploiting the computational resources available using specialized frameworks such as Message Passing Interface (MPI) or heterogeneous systems including General-purpose Processing on Graphics Processing Units (GPGPU). With the recent development of extremely fast broadband networks, it has become possible to distribute the calculations at increasing scales over different geographical locations (Schadt *et al.*, 2010). Cluster, grid and emerging cloud computing are all examples of shared computing resources where probe design algorithms must be deployed if specificity tests and alignments are to be performed with reasonable data processing times (Gardner *et al.*, 2006; Thorsen *et al.*, 2007).

Microarray formats

As mentioned above, explorative design strategies that allow the detection of unknown sequences involve the use of degenerate probes (Bontemps *et al.*, 2005; Militon *et al.*, 2007; Terrat *et al.*, 2010; Dugat-Bony *et al.*, 2011; Parisot *et al.*, 2012). The selected strategy will therefore greatly influence the choice between the two major DNA microarray types (*ex situ* or *in situ*), the platform and the density (Dufva, 2005; Ehrenreich, 2006; Kawasaki, 2006). When *in situ* synthesis microarrays such as the Agilent, Affymetrix and NimbleGen platforms are used, all non–degenerate combinations that result from a degenerate probe have to be independently synthesized. Consequently, the final number of probes (i.e. density) will exponentially increase for the array production. For instance, for the CODEHMOP (Bontemps *et al.*, 2005) and Metabolic Design (Terrat *et al.*, 2010) strategies that were developed for the FGAs probe design, the degenerate probes are derived from the multiple sequence alignment of all possible nucleotide sequences that are able to code for the targeted conserved amino acid motif. Because the genetic code often involves degeneracy at the third position of each codon, a 24-mer probe (i.e. targeting a seven amino acid conserved motif) will generate at least 128 combinations (assuming a minimal degeneracy rate of 2 for each codon). This value will reach at least 131,072 for a 51-mer probe containing 17 degenerate positions. Conversely, *ex situ* platforms allow the degenerate probes (all combinations mixed together) to be spotted in the same location on the array and consequently reduce the total number of features. However, in this latter case, the sensitivity may be affected by the complexity of the mixed oligonucleotides.

Other user choices may also affect the final number of probes per array. Replication is crucial for achieving reliable data for microarrays (Spruill *et al.*, 2002). Multiple replicates of the same probe provide some backup if a feature cannot be evaluated because of technical artefacts such as dye precipitation or dust particles. A statistical estimation has deduced that at least three replicates should be located (Lee *et al.*, 2000). Additionally, multiple probes could be designed per gene to increase confidence in the results (Loy *et al.*, 2002; Chou *et al.*, 2004a) and to mask misleading signal variations whose causes (e.g. target secondary structure, probe folding) are not yet fully understood (Pozhitkov *et al.*, 2007). Third, some platforms such as Affymetrix GeneChips determine probe pairs where each probe ('match') is accompanied by a negative control with a single differing base in a central position ('mismatch probe') to discriminate between true signals and those arising from non-specific hybridization (Lipshutz *et al.*, 1999).

To address the problem of the number of probes, several commercial companies have proposed two major types of high-density microarrays: (i) *in situ* synthesized microarrays, which are distributed by Agilent (http://www.chem.agilent.com), NimbleGen (http://www.nimblegen.com) and Affymetrix (http://www.affymetrix.com), can contain billions of probes and can be physically divided into multiple arrays per slide (up to 12) to perform simultaneous analyses of several samples in a single experiment; and (ii) spotted microarrays [e.g. Arrayit (http://www.arrayit.com)] with a current printing capacity close to 100,000 features per microarray.

Data analysis

Because probe design software programs are numerous, microarray formats are heterogeneous and biological questions are different, the analysis of microarray data results can be a major challenge. For instance, if an explorative design strategy has been performed, the signals encountered for these probes must be carefully interpreted. Consequently, a future direction for this field could be to develop automatic procedures to analyse microarray data.

Other applications

Probe design is now a common task in molecular ecology. Probes can be used in several molecular techniques including microarrays as well as PCR, quantitative PCR, and FISH. In addition, a promising strategy for reducing the complexity of environmental samples by enriching the desired genomic target using probes before sequencing is being adapted for microbial ecology. The more efficient methods rely on the complementary hybridization of nucleic acid capture probes to the targeted DNA sequences; these methods use either solid phase hybridization (e.g. using capture arrays) (Albert *et al.*, 2007; Okou *et al.*, 2007; Mokry *et al.*, 2010), or solution phase, also known as Solution Hybridization Selection (SHS) (Gnirke *et al.*, 2009; Tewhey *et al.*, 2009; Denonfoux *et al.*, 2013).

The use of explorative probe design in sequence-capture approaches that couple with next generation sequencing (NGS), such as those originally developed for the direct selection of human genomic loci (Albert *et al.*, 2007), could also improve characterization of microbial communities in microbial ecology. In fact, sequence capture elution products should allow the full identification and characterization of new taxa when using phylogenetic probes or new protein-coding genes with functional probes. Furthermore, the innovative approach developed by (Denonfoux *et al.*, 2013) aims to capture large DNA fragments and allows the identification of genes flanking the targeted biomarkers. Probe design software programs remain a popular research topic and have a promising future.

Conclusions

With the availability of high-density custom microarrays, the selection of high-quality oligonucleotide probes is a crucial task. Although microarrays are particularly well suited for the detection and quantification of genes or transcripts, accurate measurements depend on good probe design. Oligonucleotide design is an optimization task and must take into account various parameters that influence the interaction between the probe and the target. Increasingly, the recent development of computational methods as well as the increase in the number of available sequences in databases allows the selection of large probe sets with a wide spectrum of thermodynamic properties. Several software solutions are available to help the user and solve the current bottlenecks in the choice of high-quality probe sets that must combine basic criteria such as sensibility, specificity and uniformity. Each software program has advantages and drawbacks, and the choice of programmes must be made in total accordance with the nature of projects and the basic scientific question. Probe design strategies have evolved and hence become more easily computed over time, thus providing a foundation for more sophisticated work in bioinformatics. Although much progress has been achieved, probe design for microarray experiments remains a challenging and active research field.

Online resources for oligonucleotide probe design programs and databases

Oligonucleotide probe design software

ARB http://www.arb-home.de/. Probe design software dedicated to phylogenetic oligonucleotide arrays and based on the ARB-Silva database.

ArrayOligoSelector http://arrayoligosel.sourceforge.net. ArrayOligoSelector is dedicated to the design of gene specific long oligonucleotide probes for entire genomes.

CaSSiS http://cassis.in.tum.de. CaSSiS is a fast and scalable software for computing comprehensive collections of sequence- and sequence-group-specific oligonucleotide signatures from large sets of hierarchically clustered nucleic acid sequence data.

ChipD http://chipd.uwbacter.org/. The chipD Server is a tool for designing whole-genome tiling microarrays.

CommOligo http://ieg.ou.edu/software.htm. CommOligo is able to select group-specific oligonucleotide probes for functional gene arrays.

DEODAS http://deodas.sourceforge.net/. DEODAS designs consensus-degenerate oligonucleotide probes for microarrays targeting functional genes.

GoArrays http://g2im.u-clermont1.fr/serimour/goarrays.html. GoArrays' strategy allows concatenation of two short oligonucleotide probes with a random linker to generate efficient longer probes.

HiSpOD http://g2im.u-clermont1.fr/hispod. HiSpOD is a software dedicated for functional gene arrays dedicated to microbial ecology and environmental studies.

KASpOD http://g2im.u-clermont1.fr/kaspod. KASpOD is a program for designing signature sequences for various applications including in phylogenetic or functional microarray experiments.

MAMMOT http://www.mammot.org.uk/. MAMMOT is a genomic tiling array probe design and visualization tool.

Metabolic Design ftp://195.221.123.90/. Metabolic Design is a probe design software for explorative functional microarrays dedicated to microbial ecology and environmental studies.

MOPeD http://moped.genetics.emory.edu/newdesign.html. The MoPeD web-service allows designing whole-genome tiling microarrays.

MPrime http://kbrin.a-bldg.louisville.edu/Tools/OligoDesign/MPrime.html. MPrime allows the efficient selection of multiple oligonucleotides for whole-genome ORF arrays in either the human, mouse, or rat genomes.

Mprobe http://www.biosun.org.cn/mprobe/. Mprobe is software to design optimal oligonucleotides for whole-genome and functional gene microarrays.

OligoArray http://berry.engin.umich.edu/oligoarray2_1/. OligoArray computes gene specific oligonucleotides for genome-scale or functional gene oligonucleotide microarrays.

OligoPicker http://pga.mgh.harvard.edu/oligopicker/. OligoPicker is dedicated to the design of gene specific oligonucleotide probes for entire genomes.

OligoTiler http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi. The OligoTiler web-service is a tool for designing whole-genome tiling microarrays.

OligoWiz http://www.cbs.dtu.dk/services/OligoWiz2. OligoWiz is software to select optimal oligonucleotides for whole-genome and functional gene microarrays.

PhylArray http://g2im.u-clermont1.fr/serimour/phylarray. Phylarray is a program for designing oligonucleotide probes from 16S rRNA sequences for use in phylogenetic microarray experiments.

PICKY http://www.complex.iastate.edu/download/Picky/index.html. PICKY is a probe design software for selecting gene specific oligonucleotide probes based on a given gene set.

PRIMEGENS http://primegens.org/. PRIMEGENS designs gene specific oligonucleotides for genome-scale or functional gene oligonucleotide microarrays.

ProbeSelect http://stormo.wustl.edu/src/probeselect-src.tar. ProbeSelect allows the selection of multiple oligonucleotides for whole-genome ORF arrays

ProDesign http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html. ProDesign is able to select group-specific oligonucleotide probes for functional gene arrays.

Teolenn http://transcriptome.ens.fr/teolenn/. Teolenn is a universal probe design workflow for whole-genome arrays and developed with a flexible and customizable module organization.

UPS 2.0 http://array.iis.sinica.edu.tw/ups/. UPS 2.0 is software to design optimal oligonucleotides for whole-genome and functional gene microarrays.

Databases

DDBJ http://www.ddbj.nig.ac.jp/. DDBJ is the DNA DataBank of Japan, one of the three major public sequence database.

ENA www.ebi.ac.uk/ena/. The European Nucleotide Archive is one of the three major public sequence database.

Genbank http://www.ncbi.nlm.nih.gov/genbank. Genbank is the National Institutes of Health (USA) sequence database, one of the three major public sequence database.

Greengenes http://greengenes.lbl.gov. The Greengenes web application provides access to the current and comprehensive 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading.

RDP http://rdp.cme.msu.edu/. Ribosomal Database Project provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

SILVA http://www.arb-silva.de/. SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).

Uniprot (Swiss-Prot & TrEMBL) http://www.ebi.ac.uk/uniprot/. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

Other software

Bellerophon http://comp-bio.anu.edu.au/Bellerophon/. Bellerophon is a program for detecting chimeric sequences in a multiple sequence dataset by comparative analysis.

BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences.

BLASTCLUST http://toolkit.tuebingen.mpg.de/blastclust. BLASTCLUST is provided through the BLAST package and allows clustering a set of unaligned FASTA sequences by single-linkage clustering.

ClustalW http://www.ebi.ac.uk/Tools/msa/clustalw2/. ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

GBrowse http://gmod.org/wiki/GBrowse. GBrowse is a combination of database and interactive web pages for manipulating and displaying annotations on genomes.

GPU-BLAST http://eudoxus.cheme.cmu.edu/gpublast/gpublast.html. GPU-BLAST is an accelerated version of the popular BLAST using a general-purpose graphics processing unit (GPU).

mfold http://mfold.rna.albany.edu/?q=mfold. mfold package predicts secondary structures for RNA and DNA using nearest neighbour thermodynamic rules.

mpiBLAST http://www.mpiblast.org/. mpiBLAST is a parallel implementation of the popular BLAST using message passing interface (MPI)

NAST http://greengenes.lbl.gov/NAST/. NAST aligns a batch of sequences against the 16S greengenes rRNA gene database.

PatMan https://bioinf.eva.mpg.de/patman/. PatMan is a DNA pattern matcher for short sequences.

RepeatMasker http://www.repeatmasker.org/. RepeatMasker screens DNA sequences for interspersed repeats and low complexity DNA sequences.

References

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., *et al.* (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods *4*, 903–905.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Ashelford, K.E., Weightman, A.J., and Fry, J.C. (2002). PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. Nucleic Acids Res. *30*, 3481–3489.

Auer, H., Lyianarachchi, S., Newsom, D., Klisovic, M.I., Marcucci, G., Marcucci, U., and Kornacker, K. (2003). Chipping away at the chip bias: RNA degradation in microarray analysis. Nat. Genet. *35*, 292–293.

Bader, K.C., Grothoff, C., and Meier, H. (2011). Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. Bioinformatics *27*, 1546–1554.

Belosludtsev, Y.Y., Bowerman, D., Weil, R., Marthandan, N., Balog, R., Luebke, K., Lawson, J., Johnston, S.A., Lyons, C.R., Obrien, K., *et al.* (2004). Organism identification using a genome sequence-independent universal microarray probe set. BioTechniques *37*, 654–8–660.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., and Sayers, E.W. (2012). GenBank. Nucleic Acids Res. *40*, D48–D53.

Bertone, P., Gerstein, M., and Snyder, M. (2005). Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res. *13*, 259–274.

Bertone, P., Trifonov, V., Rozowsky, J.S., Schubert, F., Emanuelsson, O., Karro, J., Kao, M.-Y., Snyder, M., and Gerstein, M. (2006). Design optimization methods for genomic DNA tiling arrays. Genome Res. *16*, 271–281.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003). Development and validation of a diagnostic microbial microarray for methanotrophs. Environ. Microbiol. *5*, 566–582.

Bontemps, C., Golfier, G., Gris-Liebe, C., Carrere, S., Talini, L., and Boivin-Masson, C. (2005). Microarray-based detection and typing of the Rhizobium nodulation gene nodC: potential of DNA arrays to diagnose biological functions of interest. Appl. Environ. Microbiol. *71*, 8042–8048.

Boyce, R., Chilana, P., and Rose, T.M. (2009). iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Res. *37*, W222–W228.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B., and DeRisi, J.L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol. *4*, R9.

Brodie, E.L., DeSantis, T.Z., Joyner, D.C., Baek, S.M., Larsen, J.T., Andersen, G.L., Hazen, T.C., Richardson, P.M., Herman, D.J., Tokunaga, T.K., *et al.* (2006). Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. *72*, 6288–6298.

Brodie, E.L., DeSantis, T.Z., Parker, J.P.M., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. P. Natl. Acad. Sci. USA *104*, 299–304.

Burkhardt, S., and Kärkkäinen, J. (2003). Fast Lightweight Suffix Array Construction and Checking. CPM 2003, Lect. Notes Comput. Sc. *2676*, 55–69.

Busti, E., Bordoni, R., Castiglioni, B., Monciardini, P., Sosio, M., Donadio, S., Consolandi, C., Rossi Bernardi, L., Battaglia, C., and De Bellis, G. (2002). Bacterial discrimination by means of a universal array approach mediated by LDR (ligase detection reaction). BMC Microbiol. *2*, 27.

Candela, M., Consolandi, C., Severgnini, M., Biagi, E., Castiglioni, B., Vitali, B., De Bellis, G., and Brigidi, P. (2010). High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction – universal array approach. BMC Microbiol. *10*, 116.

Castiglioni, B., Rizzi, E., Frosini, A., Sivonen, K., Rajaniemi, P., Rantala, A., Mugnai, M.A., Ventura, S., Wilmotte, A., Boutte, C., *et al.* (2004). Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. Appl. Environ. Microbiol. *70*, 7161–7172.

Chandler, D.P., and Jarrell, A.E. (2005). Taking arrays from the lab to the field: trying to make sense of the unknown. BioTechniques *38*, 591–600.

Chandler, D.P., Newton, G.J., Small, J.A., and Daly, D.S. (2003). Sequence versus structure for the direct detection of 16S rRNA on planar oligonucleotide microarrays. Appl. Environ. Microbiol. *69*, 2950–2958.

Chen, S.-H., Lo, C.-Z., Su, S.-Y., Kuo, B.-H., Hsiung, C.A., and Lin, C.-Y. (2010). UPS 2.0: unique probe selector for probe design and oligonucleotide microarrays at the pangenomic/genomic level. BMC Genomics *11 Suppl 4*, S6.

Chou, C.-C., Chen, C.-H., Lee, T.-T., and Peck, K. (2004a). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Res. *32*, e99.

Chou, H.-H., Hsia, A.-P., Mooney, D.L., and Schnable, P.S. (2004b). Picky: oligo microarray design for large genomes. Bioinformatics *20*, 2893–2902.

Chung, W.-H., Rhee, S.-K., Wan, X.-F., Bae, J.-W., Quan, Z.-X., and Park, Y.-H. (2005). Design of long oligonucleotide probes for functional gene detection in a microbial community. Bioinformatics *21*, 4092–4100.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., *et al.* (2009). Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res. *37*, D19–D25.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. *37*, D141–D145.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002). Estimating prokaryotic diversity and its limits. P. Natl. Acad. Sci. USA *99*, 10494–10499.

Curtis, T.P., Head, I.M., Lunn, M., Woodcock, S., Schloss, P.D., and Sloan, W.T. (2006). What is the extent of prokaryotic diversity? Phil. Trans. R. Soc. B *361*, 2023–2037.

Darling, A., Carey, L., and Feng, W. (2003). The design, implementation, and evaluation of mpiBLAST. 4th International Conference on Linux Clusters and ClusterWorld *2003*.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., Hirsch, P.R., and Vogel, T.M. (2011). Accessing the soil metagenome for studies of microbial diversity. Appl. Environ. Microbiol. *77*, 1315–1324.

Denonfoux, J., Parisot, N., Dugat-Bony, E., Biderre-Petit, C., Boucher, D., Morgavi, D.P., Le Paslier, D., Peyretaillade, E., and Peyret, P. (2013). Gene Capture Coupled to High-Throughput Sequencing as a Strategy for Targeted Metagenome Exploration. DNA Res. doi:10.1093/dnares/dst001

DeSantis, T.Z., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., and Andersen, G.L. (2006a). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. *34*, W394–W399.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006b). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Andersen, G.L. (2007). High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. *53*, 371–383.

van Doorn, R., Slawiak, M., Szemes, M., Dullemans, A.M., Bonants, P., Kowalchuk, G.A., and Schoen, C.D. (2009). Robust detection and identification of multiple oomycetes and fungi in environmental samples by using a novel cleavable padlock probe-based ligation detection assay. Appl. Environ. Microbiol. *75*, 4185–4193.

Dufour, Y.S., Wesenberg, G.E., Tritt, A.J., Glasner, J.D., Perna, N.T., Mitchell, J.C., and Donohue, T.J. (2010). chipD: a web tool to design oligonucleotide probes for high-density tiling arrays. Nucleic Acids Res. *38*, W321–W325.

Dufva, M. (2005). Fabrication of high quality microarrays. Biomol. Eng. *22*, 173–184.

Dugat-Bony, E., Missaoui, M., Peyretaillade, E., Biderre-Petit, C., Bouzid, O., Gouinaud, C., Hill, D.R.C., and Peyret, P. (2011). HiSpOD: probe design for functional DNA microarrays. Bioinformatics *27*, 641–648.

Dugat-Bony, E., Biderre-Petit, C., Jaziri, F., David, M.M., Denonfoux, J., Lyon, D.Y., Richard, J.-Y., Curvers, C., Boucher, D., Vogel, T.M., *et al.* (2012a). In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. Microb. Biotechnol. *5*, 642–653.

Dugat-Bony, E., Peyretaillade, E., Parisot, N., Biderre-Petit, C., Jaziri, F., Hill, D.R.C., Rimour, S., and Peyret, P. (2012b). Detecting unknown sequences with DNA microarrays: explorative probe design strategies. Environ. Microbiol. *14*, 356–371.

Ehrenreich, A. (2006). DNA microarray technology for the microbiologist: an overview. Appl. Microbiol. Biotechnol. *73*, 255–273.

Feng, S., and Tillier, E.R.M. (2007). A fast and flexible approach to oligonucleotide probe design for genomes and gene families. Bioinformatics *23*, 1195–1202.

Franke-Whittle, I.H., Goberna, M., Pfister, V., and Insam, H. (2009). Design and development of the ANAEROCHIP microarray for investigation of methanogenic communities. J. Microbiol. Methods *79*, 279–288.

Fredrickson, H.L., Perkins, E.J., Bridges, T.S., Tonucci, R.J., Fleming, J.K., Nagel, A., Diedrich, K., Mendez-Tenorio, A., Doktycz, M.J., and Beattie, K.L. (2001). Towards environmental toxicogenomics – development of a flow-through, high-density DNA hybridization array and its application to ecotoxicity assessment. Sci. Total Environ. *274*, 137–149.

Gans, J., Wolinsky, M., and Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science *309*, 1387–1390.

Gardner, M., Feng, W.-C., Archuleta, J., Lin, H., and Ma, X. (2006). Parallel Genomic Sequence-Searching on an Ad-Hoc Grid: Experiences, Lessons Learned, and Implications. SC 2006 Conference, Proceedings of the ACM/IEEE, pp. 22–22.

Gentry, T.J., Wickham, G.S., Schadt, C.W., He, Z., and Zhou, J. (2006). Microarray applications in microbial ecology research. Microb. Ecol. *52*, 159–175.

Gerry, N.P., Witowski, N.E., Day, J., Hammer, R.P., Barany, G., and Barany, F. (1999). Universal DNA microarray method for multiplex detection of low abundance point mutations. J. Mol. Biol. *292*, 251–262.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., *et al.* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182–189.

Gräf, S., Nielsen, F.G.G., Kurtz, S., Huynen, M.A., Birney, E., Stunnenberg, H., and Flicek, P. (2007). Optimized design and assessment of whole genome tiling arrays. Bioinformatics *23*, i195–i204.

Guschin, D.Y., Mobarry, B.K., Proudnikov, D., Stahl, D.A., Rittmann, B.E., and Mirzabekov, A.D. (1997). Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl. Environ. Microbiol. *63*, 2397–2402.

Hancock, J.M., and Armstrong, J.S. (1994). SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput. Appl. Biosci. *10*, 67–70.

Handley, K.M., Wrighton, K.C., Piceno, Y.M., Andersen, G.L., DeSantis, T.Z., Williams, K.H., Wilkins, M.J., N'guessan, A.L., Peacock, A., Bargar, J., *et al.* (2012). High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment. FEMS Microbiol. Ecol. *81*, 188–204.

Hazen, T.C., Dubinsky, E.A., DeSantis, T.Z., Andersen, G.L., Piceno, Y.M., Singh, N., Jansson, J.K., Probst, A., Borglin, S.E., Fortney, J.L., *et al.* (2010). Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science *330*, 204–208.

He, Z., Wu, L., Li, X., Fields, M.W., and Zhou, J. (2005). Empirical establishment of oligonucleotide probe design criteria. Appl. Environ. Microbiol. *71*, 3753–3760.

He, Z., Gentry, T.J., Schadt, C.W., Wu, L., Liebich, J., Chong, S.C., Huang, Z., Wu, W., Gu, B., Jardine, P., *et al.* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J. *1*, 67–77.

He, Z., Van Nostrand, J.D., Wu, L., and Zhou, J. (2008). Development and application of functional gene arrays for microbial community analysis. Trans. Nonferrous Met. Soc. China *18*, 1319–1327.

He, Z., Deng, Y., Van Nostrand, J.D., Tu, Q., Xu, M., Hemme, C.L., Li, X., Wu, L., Gentry, T.J., Yin, Y., *et al.* (2010). GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. ISME J. *4*, 1167–1179.

He, Z., Van Nostrand, J.D., Deng, Y., and Zhou, J. (2011). Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. Front. Environ. Sci. Engin. China. *5*, 1–20.

He, Z., Deng, Y., and Zhou, J. (2012a). Development of functional gene microarrays for microbial community analysis. Curr. Opin. Biotechnol. *23*, 49–55.

He, Z., Van Nostrand, J.D., and Zhou, J. (2012b). Applications of functional gene microarrays for profiling microbial communities. Curr. Opin. Biotechnol. *23*, 460–466.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics *20*, 2317–2319.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., *et al.* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. *19*, 342–347.

Hultman, J., Ritari, J., Romantschuk, M., Paulin, L., and Auvinen, P. (2008). Universal ligation-detection-reaction microarray applied for compost microbes. BMC Microbiol. *8*, 237.

Huyghe, A., François, P., Charbonnier, Y., Tangomo-Bento, M., Bonetti, E.-J., Paster, B.J., Bolivar, I., Baratti-Mayer, D., Pittet, D., Schrenzel, J., *et al.* (2008). Novel microarray design strategy to study complex bacterial communities. Appl. Environ. Microbiol. *74*, 1876–1885.

Jacobs, K.A., Rudersdorf, R., Neill, S.D., Dougherty, J.P., Brown, E.L., and Fritsch, E.F. (1988). The thermal stability of oligonucleotide duplexes is sequence independent in tetraalkylammonium salt solutions: application to identifying recombinant DNA clones. Nucleic Acids Res. *16*, 4637–4650.

Jourdren, L., Duclos, A., Brion, C., Portnoy, T., Mathis, H., Margeot, A., and Le Crom, S. (2010). Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments. Nucleic Acids Res. *38*, e117.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Kaderali, L., and Schliep, A. (2002). Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics *18*, 1340–1349.

Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K., *et al.* (2011). DDBJ progress report. Nucleic Acids Res. *39*, D22–D27.

Kämpke, T., Kieninger, M., and Mecklenburg, M. (2001). Efficient primer design algorithms. Bioinformatics *17*, 214–225.

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D., and Madore, S.J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. *28*, 4552–4557.

Kang, S., Denman, S.E., Morrison, M., Yu, Z., Dore, J., Leclerc, M., and McSweeney, C.S. (2010). Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. Inflamm. Bowel Dis. *16*, 2034–2042.

Kaplinski, L., Scheler, O., Parkel, S., Palta, P., Toome, K., Kurg, A., and Remm, M. (2010). Detection of tmRNA molecules on microarrays at low temperatures using helper oligonucleotides. BMC Biotechnol. *10*, 34.

Kawasaki, E.S. (2006). The end of the microarray Tower of Babel: will universal standards lead the way? J. Biomol. Tech. *17*, 200–206.

Koltai, H., and Weingarten-Baror, C. (2008). Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. Nucleic Acids Res. *36*, 2395–2405.

Kostić, T., Weilharter, A., Rubino, S., Delogu, G., Uzzau, S., Rudi, K., Sessitsch, A., and Bodrossy, L. (2007). A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. Anal. Biochem. *360*, 244–254.

Lee, M.L., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. P. Natl. Acad. Sci. USA *97*, 9834–9839.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., *et al.* (2011). The European Nucleotide Archive. Nucleic Acids Res. *39*, D28–D31.

Lemoine, S., Combes, F., and Le Crom, S. (2009). An evaluation of custom microarray applications: the oligonucleotide design challenge. Nucleic Acids Res. *37*, 1726–1739.

Leparc, G.G., Tüchler, T., Striedner, G., Bayer, K., Sykacek, P., Hofacker, I.L., and Kreil, D.P. (2009). Model-based probe set optimization for high-performance microarrays. Nucleic Acids Res. *37*, e18.

Leski, T.A., Lin, B., Malanoski, A.P., and Stenger, D.A. (2012). Application of resequencing microarrays in microbial detection and characterization. Future Microbiol. *7*, 625–637.

Letowski, J., Brousseau, R., and Masson, L. (2004). Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J. Microbiol. Methods *57*, 269–278.

Li, F., and Stormo, G.D. (2001). Selection of optimal DNA oligos for gene expression arrays. Bioinformatics *17*, 1067–1076.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, W., and Ying, X. (2006). Mprobe 2.0: computer-aided probe design for oligonucleotide microarray. Appl. Bioinformatics *5*, 181–186.

Li, X., He, Z., and Zhou, J. (2005). Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res. *33*, 6114–6123.

Liebich, J., Schadt, C.W., Chong, S.C., He, Z., Rhee, S.-K., and Zhou, J. (2006). Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Appl. Environ. Microbiol. *72*, 1688–1691.

Liles, M.R., Turkmen, O., Manske, B.F., Zhang, M., Rouillard, J.-M., George, I., Balser, T., Billor, N., and Goodman, R.M. (2010). A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. Soil Biol. Biochem. *42*, 739–747.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. Nat. Genet. *21*, 20–24.

Loy, A., and Bodrossy, L. (2006). Highly parallel microbial diagnostics using oligonucleotide microarrays. Clin. Chim. Acta *363*, 106–119.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., Schleifer, K.-H., and Wagner, M. (2002). Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl. Environ. Microbiol. *68*, 5064–5081.

Loy, A., Schulz, C., Lücker, S., Schöpfer-Wendels, A., Stoecker, K., Baranyi, C., Lehner, A., and Wagner, M. (2005). 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order 'Rhodocyclales'. Appl. Environ. Microbiol. *71*, 1373–1386.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., *et al.* (2004). ARB: a software environment for sequence data. Nucleic Acids Res. *32*, 1363–1371.

Maldonado-Rodriguez, R., Espinosa-Lara, M., Loyola-Abitia, P., Beattie, W.G., and Beattie, K.L. (1999). Mutation detection by stacking hybridization on genosensor arrays. Mol. Biotechnol. *11*, 13–25.

Manber, U., and Myers, G. (1993). Suffix arrays: a new method for on-line string searches. SIAM Journal on Computing *22*.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618.

Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F.C., Shen, M.-M., Lu, G., Fang, J., Liu, W.-M., Ryder, T., *et al.* (2003). Probe selection for high-density oligonucleotide arrays. P. Natl. Acad. Sci. U.S.A. *100*, 11237–11242.

Meng, D., Broschat, S.L., and Call, D.R. (2008). A Java-based tool for the design of classification microarrays. BMC Bioinformatics *9*, 328.

Militon, C., Rimour, S., Missaoui, M., Biderre-Petit, C., Barra, V., Hill, D.R.C., Moné, A., Gagne, G., Meier, H., Peyretaillade, E., *et al.* (2007). PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics *23*, 2550–2557.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics *85*, 1–15.

Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V., and Cuppen, E. (2010). Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res. *38*, e116.

Myers, G. (1999). A fast bit-vector algorithm for approximate string matching based on dynamic programming. J. ACM *46*, 395–415.

Nakaya, H., Reis, E., and Verjovski-Almeida, S. (2007). Concepts on Microarray Design for Genome and Transcriptome Analyses. In Nucleic Acids Hybridization Modern Applications, A. Buzdin, and S. Lukyanov, eds. (Springer Netherlands), pp. 265–307.

Neufeld, J.D., Mohn, W.W., and de Lorenzo, V. (2006). Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. Environ. Microbiol. *8*, 126–140.

Nordberg, E.K. (2005). YODA: selecting signature oligonucleotides. Bioinformatics *21*, 1365–1370.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., and Zwick, M.E. (2007). Microarray-based genomic selection for high-throughput resequencing. Nat. Methods *4*, 907–909.

Paliy, O., and Agans, R. (2012). Application of phylogenetic microarrays to interrogation of human microbiota. FEMS Microbiol. Ecol. *79*, 2–11.

Palmer, C., Bik, E.M., Eisen, M.B., Eckburg, P.B., Sana, T.R., Wolber, P.K., Relman, D.A., and Brown, P.O. (2006). Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. *34*, e5.

Pariset, L., Chillemi, G., Bongiorni, S., Romano Spica, V., and Valentini, A. (2009). Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. Nat. Biotechnol. *25*, 272–279.

Parisot, N., Denonfoux, J., Dugat-Bony, E., Peyret, P., and Peyretaillade, E. (2012). KASpOD – a web service for highly specific and explorative oligonucleotide design. Bioinformatics *28*, 3161–3162.

Patel, V.C., Mondal, K., Shetty, A.C., Horner, V.L., Bedoyan, J.K., Martin, D., Caspary, T., Cutler, D.J., and Zwick, M.E. (2010). Microarray oligonucleotide probe designer (MOPeD): A web service. Open Access Bioinformatics *2*, 145–155.

Peplies, J., Glöckner, F.O., and Amann, R. (2003). Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl. Environ. Microbiol. *69*, 1397–1407.

Phillippy, A.M., Deng, X., Zhang, W., and Salzberg, S.L. (2009). Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics *10*, 293.

Pozhitkov, A.E., and Tautz, D. (2002). An algorithm and program for finding sequence specific oligonucleotide probes for species identification. BMC Bioinformatics *3*, 9.

Pozhitkov, A.E., Tautz, D., and Noble, P.A. (2007). Oligonucleotide microarrays: widely applied – poorly understood. Brief. Funct. Genomic Proteomic *6*, 141–148.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. *35*, 7188–7196.

Prüfer, K., Stenzel, U., Dannemann, M., Green, R.E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. Bioinformatics *24*, 1530–1531.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008). The rational exploration of microbial diversity. ISME J. *2*, 997–1006.

Rajilić-Stojanović, M., Heilig, H.G.H.J., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., and de Vos, W.M. (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ. Microbiol. *11*, 1736–1751.

Relógio, A., Schwager, C., Richter, A., Ansorge, W., and Valcárcel, J. (2002). Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res. *30*, e51.

Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G., and Fayard, J.-M. (2004). ROSO: optimizing oligonucleotide probes for microarrays. Bioinformatics *20*, 271–273.

Rimour, S., Hill, D.R.C., Militon, C., and Peyret, P. (2005). GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics *21*, 1094–1103.

Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., and Henikoff, S. (1998). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res. *26*, 1628–1635.

Rose, T.M., Henikoff, J.G., and Henikoff, S. (2003). CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res. *31*, 3763–3766.

Rouchka, E.C., Khalyfa, A., and Cooper, N.G.F. (2005). MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. BMC Bioinformatics *6*, 175.

Rouillard, J.-M., Zuker, M., and Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. *31*, 3057–3062.

Rychlik, W., Spencer, W.J., and Rhoads, R.E. (1990). Optimization of the annealing temperature for DNA amplification *in vitro*. Nucleic Acids Res. *18*, 6409–6412.

Ryder, E., Jackson, R., Ferguson-Smith, A., and Russell, S. (2006). MAMMOT – a set of tools for the design, management and visualization of genomic tiling arrays. Bioinformatics *22*, 883–884.

Sadakane, K., and Shibuya, T. (2001). Indexing huge genome sequences for solving various problems. Genome Inform. *12*, 175–183.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor

thermodynamics. Proc. Natl. Acad. Sci. U.S.A. *95*, 1460–1465.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., and Nolan, G.P. (2010). Computational solutions to large-scale data management and analysis. Nat. Rev. Genet. *11*, 647–657.

Schliep, A., and Rahmann, S. (2006). Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree. Bioinformatics *22*, e424–e430.

Schliep, A., and Krause, R. (2008). Efficient algorithms for the computational design of optimal tiling arrays. IEEE ACM T. Comput. Bi. *5*, 557–567.

Schönmann, S., Loy, A., Wimmersberger, C., Sobek, J., Aquino, C., Vandamme, P., Frey, B., Rehrauer, H., and Eberl, L. (2009). 16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genus *Burkholderia*. Environ. Microbiol. *11*, 779–800.

Selinger, D.W., Cheung, K.J., Mei, R., Johansson, E.M., Richmond, C.S., Blattner, F.R., Lockhart, D.J., and Church, G.M. (2000). RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nat. Biotechnol. *18*, 1262–1268.

Severgnini, M., Cremonesi, P., Consolandi, C., Caredda, G., De Bellis, G., and Castiglioni, B. (2009). ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. Nucleic Acids Res. *37*, e109.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., and Chandler, D.P. (2001). Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl. Environ. Microbiol. *67*, 4708–4716.

Sokal, R.R., and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. *38*, 1409–1438.

Spruill, S.E., Lu, J., Hardy, S., and Weir, B. (2002). Assessing sources of variability in microarray gene expression data. BioTechniques *33*, 916–20–922–3.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., *et al.* (2002). The generic genome browser: a building block for a model organism system database. Genome Res. *12*, 1599–1610.

Stralis-Pavese, N., Sessitsch, A., Weilharter, A., Reichenauer, T., Riesing, J., Csontos, J., Murrell, J.C., and Bodrossy, L. (2004). Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. Environ. Microbiol. *6*, 347–363.

Suzuki, M.T., and Giovannoni, S.J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. *62*, 625–630.

Szemes, M., Bonants, P., de Weerdt, M., Baner, J., Landegren, U., and Schoen, C.D. (2005). Diagnostic application of padlock probes – multiplex detection of plant pathogens using universal microarrays. Nucleic Acids Res. *33*, e70–e70.

Talla, E., Tekaia, F., Brino, L., and Dujon, B. (2003). A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. BMC Genomics *4*, 38.

Taroncher-Oldenburg, G., Griner, E.M., Francis, C.A., and Ward, B.B. (2003). Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl. Environ. Microbiol. *69*, 1159–1171.

Terrat, S., Peyretaillade, E., Goncalves, O., Dugat-Bony, E., Gravelat, F., Moné, A., Biderre-Petit, C., Boucher, D., Troquet, J., and Peyret, P. (2010). Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. BMC Bioinformatics *11*, 478.

Tewhey, R., Nakano, M., Wang, X., Pabón-Peña, C., Novak, B., Giuffre, A., Lin, E., Happe, S., Roberts, D.N., LeProust, E.M., *et al.* (2009). Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. *10*, R116.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Thorsen, O., Smith, B., Sosa, C.P., Jiang, K., Lin, H., Peters, A., and Feng, W.-C. (2007). Parallel Genomic Sequence-Search on a Massively Parallel System. (New York, USA: ACM Press), pp. 59–68.

Tijssen, P. (1993). Hybridization With Nucleic Acid Probes. Laboratory Techniques in Biochemistry and Molecular Biology, PC vand der Vliet, ed., c.

Tiquia, S.M., Wu, L., Chong, S.C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004). Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. BioTechniques *36*, 664–70–672–674–5.

Tomiuk, S., and Hofmann, K. (2001). Microarray probe selection strategies. Brief. Bioinform. *2*, 329–340.

Vijaya Satya, R., Zavaljevski, N., Kumar, K., Bode, E., Padilla, S., Wasieloski, L., Geyer, J., and Reifman, J. (2008). In silico microarray probe design for diagnosis of multiple pathogens. BMC Genomics *9*, 496.

Vora, G.J., Meador, C.E., Stenger, D.A., and Andreadis, J.D. (2004). Nucleic acid amplification strategies for DNA microarray-based pathogen detection. Appl. Environ. Microbiol. *70*, 3047–3054.

Vouzis, P.D., and Sahinidis, N.V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. Bioinformatics *27*, 182–188.

Wagner, M., Smidt, H., Loy, A., and Zhou, J. (2007). Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microb. Ecol. *53*, 498–506.

Wang, X., and Seed, B. (2003). Selection of oligonucleotide probes for protein coding sequences. Bioinformatics *19*, 796–802.

Wernersson, R., and Nielsen, H.B. (2005). OligoWiz 2.0 – integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. *33*, W611–W615.

Wernersson, R., Juncker, A.S., and Nielsen, H.B. (2007). Probe selection for DNA microarrays using OligoWiz. Nat. Protoc. *2*, 2677–2691.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002). High-density microarray of small-subunit ribosomal DNA probes. Appl. Environ. Microbiol. *68*, 2535–2541.

Wood, W.I., Gitschier, J., Lasky, L.A., and Lawn, R.M. (1985). Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. U.S.A. *82*, 1585–1588.

Wootton, J.C., and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry *17*, 149–163.

Xu, D., Li, G., Wu, L., Zhou, J., and Xu, Y. (2002). PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics *18*, 1432–1437.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., *et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science *302*, 842–846.

Yershov, G., Barsky, V., Belgovskiy, A., Kirillov, E., Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, A., Dubiley, S., *et al.* (1996). DNA analysis and diagnostics on oligonucleotide microchips. Proc. Natl. Acad. Sci. USA *93*, 4913–4918.

Zhou, J. (2003). Microarrays for bacterial detection and microbial community analysis. Curr. Opin. Microbiol. *6*, 288–294.

Zhou, Z., Dou, Z.-X., Zhang, C., Yu, H.-Q., Liu, Y.-J., Zhang, C.-Z., and Cao, Y.-J. (2007). A strategy to optimize the oligo-probes for microarray-based detection of viruses. Virol. Sin. *22*, 326–335.

Ziv, J., and Lempel, A. (1977). A universal algorithm for sequential data compression. IEEE Trans. Inform. Theory *23*, 337–343.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Bien qu'illustrés au travers des biopuces ADN, les concepts généraux et les différentes stratégies de détermination des sondes peuvent être appliqués à d'autres méthodes d'études des microorganismes comme la capture de gènes. En effet, ces stratégies de détermination de sondes sont le fruit d'une réflexion portant sur de nombreux paramètres. Il est possible de mettre en avant certains d'entre eux comme la longueur ayant un impact direct sur la spécificité et la sensibilité de la détection des cibles. Ainsi, la discrimination phylogénétique de microorganismes sur la base de leurs séquences ADNr 16S (marqueur montrant de fortes similarités de séquences) se fera avec des sondes courtes plus spécifiques, alors que des sondes longues beaucoup plus sensibles pourront être utilisées pour l'identification de gènes codant des protéines impliquées dans des voies métaboliques au sein d'un environnement complexe. Dans tous les cas, une attention particulière doit être apportée à la sélection des sondes et donc aux logiciels utilisés (Dugat-Bony *et al.* 2012b).

Enfin, les logiciels doivent être suffisamment performants pour réaliser tous les tests de sélection des sondes avec des masses de données en constante augmentation. Ainsi, plusieurs orientations émergent avec notamment le remplacement des étapes d'alignements, la possibilité de construire des bases de données dédiées (spécifiques de l'environnement exploré) pour les tests de spécificité et l'utilisation de nouveaux moyens de calcul. Ces derniers passent par l'utilisation de machines multi-processeurs *via* les bibliothèques de fonctions « *Message Passing Interface* » (MPI), ou les processeurs des cartes graphiques (*General-purpose Processing on Graphics Processing Units*, GPGPU). De même, il est possible désormais *via* le développement des architectures de type *clusters* ou grilles de calcul de distribuer ces calculs à large échelle et à différents endroits géographiques. Ce chapitre montre donc qu'il existe une recherche bioinformatique très active dans le domaine de la conception de sondes.

Conclusion générale

Les informations disponibles sur la structure et le fonctionnement des écosystèmes restent, à l'heure actuelle, très incomplètes du fait de l'extraordinaire diversité des communautés microbiennes. Les méthodes culturales ou encore les méthodes moléculaires générant des données parcellaires basées sur l'amplification spécifique de biomarqueurs, ont laissé place à l'essor de nouvelles stratégies d'étude globale des écosystèmes. Ainsi, le développement de la métagénomique, l'utilisation du séquençage massif, des techniques moléculaires à haut-débit comme les biopuces ADN ou les nouvelles méthodes de réduction de complexité, permettent d'explorer la diversité microbienne au sein des environnements complexes.

Ces méthodes de réduction ciblée de la complexité permettent de répondre rapidement et efficacement aux problématiques actuelles d'écologie microbienne en s'intéressant uniquement à la fraction la plus informative de l'échantillon étudié pour répondre à la question biologique posée. L'efficacité de la plupart de ces méthodes repose essentiellement sur la qualité des sondes sélectionnées en termes de sensibilité, de spécificité et d'uniformité (Loy & Bodrossy 2006 ; Wagner *et al.* 2007). La détermination de ces sondes doit alors considérer un grand nombre de paramètres et demeure donc difficile. Cette complexité est d'autant plus importante lorsqu'il s'agit de déterminer des sondes exploratoires ciblant des variants génétiques encore non référencés dans les bases de données (Dugat-Bony *et al.* 2012b). Il est donc nécessaire de faire évoluer les stratégies de détermination des sondes en intégrant ce concept. De plus, il est primordial de prendre en compte l'apport exponentiel des séquences dans les bases de données, qui de par son importance nécessite aussi de repenser les algorithmes de détermination des sondes pour réduire les temps de calcul. Les différentes approches de détermination de sondes développées au cours de cette thèse feront donc l'objet d'un chapitre complet de ce manuscrit (PARTIE 2 : Détermination de sondes oligonucléotidiques). Par la suite, les diverses applications moléculaires et informatiques des sondes oligonucléotidiques précédemment déterminées seront présentées (PARTIE 3 : Applications moléculaires et bioinformatiques).

PARTIE 2 : Détermination de sondes oligonucléotidiques

1. Amélioration et déploiement sur la grille de calculs d'un logiciel de détermination de sondes oligonucléotidiques pour biopuces phylogénétiques : PhylGrid 2.0

1.1 Contexte

L'essor spectaculaire des techniques moléculaires ces dernières années a révolutionné le domaine de l'écologie microbienne et la manière d'explorer la diversité du monde microbien. De nombreux outils, dits à haut-débit, ont ainsi été développés permettant de générer d'importantes quantités de données qui étaient auparavant inaccessibles, mais qui demeurent indispensables pour comprendre le fonctionnement des écosystèmes. Les nouvelles méthodes moléculaires à haut-débit comme les biopuces ADN ou le développement de nouvelles approches de capture de gènes couplées au séquençage haut débit, démontrent être des outils pertinents pour explorer la diversité microbienne des environnements complexes (Roh *et al.* 2010).

Cependant, l'efficacité de ces méthodes moléculaires, qui nécessitent l'utilisation d'oligonucléotides, dépend entièrement des sondes sélectionnées. Celles-ci doivent être hautement sensibles et reconnaître spécifiquement les marqueurs ciblés, même ceux présents en faible abondance dans l'échantillon étudié (Gentry *et al.* 2006). De plus, du fait du grand nombre de groupes bactériens encore méconnus, il est nécessaire que ces sondes possèdent un caractère exploratoire afin d'appréhender la totalité des communautés bactériennes : connues et encore inconnues. Seul le logiciel PhylArray (Militon *et al.* 2007) permet la détermination de telles sondes pour la construction de biopuces phylogénétiques.

La première étape de l'algorithme PhylArray repose sur l'extraction de toutes les séquences d'un taxon ciblé à partir d'une base de données de séquences d'ADNr 16S expertisée. Les séquences extraites sont alors alignées en utilisant le logiciel ClustalW (Larkin *et al.* 2007). Une séquence consensus, pouvant contenir des positions dégénérées (selon la nomenclature IUPAC (Cornish-Bowden 1985)), est ensuite déduite de cet alignement multiple en intégrant la variabilité de séquence à chaque site moléculaire. Les sondes candidates, pouvant être dégénérées, sont ensuite sélectionnées le long de la séquence consensus et comparées à la base de données de séquences d'ADNr 16S pour contrôler leur spécificité.

Lorsqu'il s'agit d'une sonde dégénérée, toutes les combinaisons non dégénérées qui en sont déduites sont alors testées.

L'application du logiciel PhylArray à la construction d'une biopuce phylogénétique exhaustive, ciblant plusieurs milliers de genres bactériens, reste néanmoins délicate. En effet, la spécificité des sondes déterminées par PhylArray est évaluée grâce à une recherche de similarités (*i.e.* BLAST) contre une base de données complète de séquences d'ADNr 16S. Néanmoins, l'augmentation croissante du nombre de séquences dans les bases de données implique des temps de calculs de plus en plus longs pouvant atteindre alors jusqu'à plusieurs mois pour assurer la détermination de sondes mais également pour évaluer leur spécificité.

1.2 Objectif

L'objectif de ce travail a donc consisté à améliorer le logiciel PhylArray pour lui permettre de traiter la masse de données génomiques actuelle.

En tirant profit des architectures de calculs de plus en plus performantes, toutes les étapes de l'algorithme PhylArray ont été modifiées. L'étape initiale d'alignement multiple des séquences du genre ciblé, qui utilisait le logiciel ClustalW (Larkin *et al.* 2007), pouvait durer plusieurs heures lorsque le genre en question était représenté par de nombreuses séquences. En effet, l'algorithme de ClustalW repose sur la détermination préalable d'une matrice de distances entre toutes les paires de séquences. Cette étape initiale a été remplacée par une version parallélisée de ClustalW : ClustalW-MPI (Li 2003), qui permet de répartir le calcul de la matrice de distances sur plusieurs processeurs en parallèle, apportant ainsi un gain de temps considérable. De même, les tests de spécificité représentent la majeure partie des temps de calculs du logiciel PhylArray, notamment lorsque la séquence consensus s'avère particulièrement dégénérée. En effet, alors qu'une sonde non dégénérée ne nécessitera qu'une seule analyse BLAST, une sonde dégénérée possédant 128 combinaisons impliquera la réalisation d'autant d'analyses BLAST pour tester sa spécificité. Cette étape a elle aussi été parallélisée pour permettre de distribuer les calculs sur plusieurs dizaines voire centaines de processeurs.

L'amélioration principale apportée à l'algorithme a donc été son déploiement sur la grille de calculs EGI (*European Grid Infrastructure*) permettant ainsi d'avoir accès à plusieurs centaines de milliers de processeurs.

1.3 Principaux résultats

Le travail réalisé a permis de proposer un nouveau logiciel de détermination de sondes oligonucléotidiques utilisant pour la première fois une grille de calculs. Ce logiciel, nommé PhylGrid 2.0 a donné lieu à une publication dans le journal « *The Scientific World Journal* ».

Les différentes modifications du logiciel PhylArray ont permis une diminution drastique des temps de calculs au sein de PhylGrid 2.0. Pour exemple, l'apport du logiciel ClustalW-MPI a permis de diminuer d'un facteur 4 le temps nécessaire à l'alignement multiple des séquences. Ainsi, en utilisant 100 processeurs, les séquences du genre *Bacillus* (3947 séquences dans la base de données utilisée) ont pu être alignées en 52 minutes contre 3,5 heures avec ClustalW utilisant un seul processeur (diminution d'un facteur 4). Il a également été démontré que l'utilisation de la grille de calculs *via* PhylGrid 2.0 permettait d'accélérer d'un facteur 100 la détermination de sondes. Ce résultat a été obtenu pour le *design* de sondes correspondant à 10 genres bactériens (*Arcanobacterium*, *Bacteriovorax*, *Campylobacter*, *Citrobacter*, *Eubacterium*, *Haemophilus*, *Kaistobacter*, *Neisseria*, *Propionibacterium* et *Riemerella*). Alors que les estimations annonçaient un temps de calculs de plus de 8 mois sur un seul processeur, la détermination de sondes a pu être effectuée en moins de 55 heures à en utilisant une méthode de répartition de charge *via* PhylGrid 2.0.

Cet outil a donc été utilisé pour la sélection à grande échelle d'un jeu de sondes pour biopuces phylogénétiques. A partir d'une base de données de 66 075 séquences, représentant 2069 genres bactériens, PhylGrid 2.0 a identifié 3 553 975 sondes candidates de 25 nucléotides. Les tests de spécificité ont ensuite été réalisés en autorisant jusqu'à deux mésappariements. Toute séquence n'appartenant pas au genre ciblé et présentant deux mésappariements ou moins a donc été considérée comme pouvant potentiellement entraîner une hybridation croisée. Les cinq meilleures sondes de chaque genre ont pu être sélectionnées suivant les critères de spécificité aboutissant au final à un jeu de 19 874 sondes oligonucléotidiques pour les 2069 genres procaryotes étudiés.

Article n°3

Large Scale Explorative Oligonucleotide Probe Selection for Thousands of Genetic Groups on a Computing Grid: Application to Phylogenetic Probe Design Using a Curated Small Subunit Ribosomal RNA Gene Database.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 350487, 9 pages
http://dx.doi.org/10.1155/2014/350487

Research Article

Large Scale Explorative Oligonucleotide Probe Selection for Thousands of Genetic Groups on a Computing Grid: Application to Phylogenetic Probe Design Using a Curated Small Subunit Ribosomal RNA Gene Database

Faouzi Jaziri,[1,2] Eric Peyretaillade,[2,3] Mohieddine Missaoui,[1,2] Nicolas Parisot,[2,4] Sébastien Cipière,[1] Jérémie Denonfoux,[2,4] Antoine Mahul,[5] Pierre Peyret,[2,3] and David R. C. Hill[1]

[1] *UMR CNRS 6158, ISIMA/LIMOS, Clermont Université et Université Blaise Pascal, F63173 Aubière, France*
[2] *Clermont Université et Université d'Auvergne, EA 4678 CIDAM, BP 10448, F63001 Clermont-Ferrand Cedex 1, France*
[3] *Clermont Université et Université d'Auvergne, UFR Pharmacie, F63001 Clermont-Ferrand Cedex 1, France*
[4] *CNRS, UMR 6023, LMGE, F63171 Aubière, France*
[5] *Clermont Université, CRRI, F63177 Aubière, France*

Correspondence should be addressed to Pierre Peyret; pierre.peyret@udamail.fr and David R. C. Hill; drch@isima.fr

Received 25 September 2013; Accepted 5 December 2013; Published 6 January 2014

Academic Editors: Y. Lai and S. Ma

Phylogenetic Oligonucleotide Arrays (POAs) were recently adapted for studying the huge microbial communities in a flexible and easy-to-use way. POA coupled with the use of explorative probes to detect the unknown part is now one of the most powerful approaches for a better understanding of microbial community functioning. However, the selection of probes remains a very difficult task. The rapid growth of environmental databases has led to an exponential increase of data to be managed for an efficient design. Consequently, the use of high performance computing facilities is mandatory. In this paper, we present an efficient parallelization method to select known and explorative oligonucleotide probes at large scale using computing grids. We implemented a software that generates and monitors thousands of jobs over the European Computing Grid Infrastructure (EGI). We also developed a new algorithm for the construction of a high-quality curated phylogenetic database to avoid erroneous design due to bad sequence affiliation. We present here the performance and statistics of our method on real biological datasets based on a phylogenetic prokaryotic database at the genus level and a complete design of about 20,000 probes for 2,069 genera of prokaryotes.

1. Introduction

The total number of species on our planet is of about 9 million, according to the latest biodiversity estimate. However, the vast majority of these species are not yet discovered and only over 1.2 million species have been already catalogued in a central database [1]. Most nondescribed species are microorganisms. Microbial communities represent the most important and diverse group of organisms living on earth. They play an important role in the functioning of ecosystems [2]. The comprehension of the role of microorganisms is then a major challenge of microbial ecology. Because of the huge microbial biocomplexity, high-throughput molecular tools allowing simultaneous analyses of existing populations are well adapted to survey microorganisms in complex environments [3].

Phylogenetic Oligonucleotide Arrays (POAs) are currently widely used and are one of the most promising approaches for studying microbial communities. They generally use oligonucleotide probes to target small subunit ribosomal RNA (SSU rRNA) genes and discriminate organisms. SSU rRNA gene is a phylogenetic biomarker largely used in the

majority of studies. However, the sequences could be highly conserved leading to some difficulties for species discrimination. Consequently, specific oligonucleotide probes selection for POAs could be a very difficult task to obtain a high resolution level [4].

Efficient oligonucleotide probes must have the following two properties: sensitive and specific. The sensitivity of a probe means its capacity to detect low levels of its complementary target in complex samples. A sensitive probe is one that is able to access its complementary sequence in the target and returns a strong signal when the target is present in the hybridized sample. The sensitivity generally increases with probe length as the binding energy for longer probe/target hybrid complexes is typically higher and hybridization kinetics are irreversible.

The specificity of a probe means its capacity to hybridize only with its complementary counterpart target. A specific probe is one that does not cross-hybridize with a nontarget sequence and returns a weak signal when the target is absent from the hybridized sample. The specificity generally decreases with the increase of probe length: short oligonucleotide probes are more specific, allowing discrimination of single nucleotide polymorphisms under optimal conditions, but at the cost of reduced sensitivity. The specificity is the most important criterion of the probes quality measure in probe design algorithms [5]. Probe design algorithms usually use specific algorithms such as suffix array method or BLAST [6] to check the specificity of probes by searching possible cross-hybridizations against datasets. However, the exponential increase of the number of sequences deposited in public databases induced an important increase in the computational capacity requirements of oligonucleotide probe design algorithms [7] and also a fundamental change in the way these algorithms are designed.

It is true that we can find fast probe design software running on regular PCs because they allow selecting probes for few DNA sequences or/and do not check the specificity of the obtained probes. The probe specificity tests against the large and ever growing biological datasets require a particular attention to develop a new generation of probe design software able to deal with high performance computing. In this context, parallel and distributed architectures such as computing clusters or computing grids [8] can provide interesting performances. Computing grids provide a promising approach to use distributed resources to meet the continuously evolving computational needs of bioinformatics tools [9]. They are particularly suited when the parallelism can be based on data splitting providing true independent computing [10]. They allow a transparent use of geographically dispersed resources for largescale distributed applications. They are adapted for time consuming algorithms that can be split into several independent jobs.

In addition to the use of known probes in POAs that allow us to simultaneously study several thousand known organisms, it is also important to design explorative probes that can detect unknown sequences not yet available in public databases and explore the vast majority of microorganisms that are still nondiscovered [3].

Here, we present a new parallelization method of a probe design algorithm to select known and explorative oligonucleotide probes using a computing grid. This software runs on the European Grid Infrastructure (EGI). EGI is a multidisciplinary grid infrastructure providing more than 250.000 CPU cores and more than 100 petabytes over 51 countries (http://www.egi.eu/). We introduced an efficient parallelization method to take advantage of the computing power available in the EGI grid to perform largescale oligonucleotides selection. We present also a new algorithm for the construction of a personal high-quality phylogenetic database that can be used to select specific, sensitive, and explorative probes targeting any prokaryotic or fungal taxonomic group, for phylogenetic oligonucleotide microarrays.

2. Related Works and Limitations

Phylogenetic Oligonucleotide Arrays (POAs), targeting the SSU rRNA genes, are known as one of the most interesting approaches to study the microbial diversity in complex environments [11]. In the last ten years, several works were done to study the biodiversity of different environments using such POAs. A microarray composed of 132 probes of length 18 mers was proposed to monitor prokaryotic microorganisms involved in sulphate reduction [12]. Another microarray considered as the most evolved POA called "PhyloChip" was developed by Brodie et al. [13] based on the Affymetrix GeneChip platform. The PhyloChip is composed of nearly 500 000 oligonucleotide probes targeting almost 9000 operational taxonomic units. This tool has been used to characterize prokaryotic communities from various ecosystems [13–17].

Additionally, several tools were proposed to select probes for phylogenetic arrays; they are discussed hereafter and in Dugat-Bony et al. [3].

The PRIMROSE program [18] was proposed to select both oligonucleotide probes and PCR primers. The probe design mechanism of PRIMROSE consists in first producing a multiple alignment for a given group of sequences. Probes are then selected and subsequently tested against an input database, to search for potential cross-hybridizations and to verify the coverage of the targeted group of sequences. PRIMROSE has been mainly used in PCR-based and FISH (fluorescent in situ hybridization) approaches [19, 20], but only a few applications of POAs using PRIMROSE have been reported [21]. The PRIMROSE tool does not allow selecting explorative probes. The ARB software package [22] proposed a probe design tool that allows selecting oligonucleotide probes with a length of 10 to 100 mers. This tool consists in searching all possible signature sequences of a targeted group of organisms specified by the user. Probes are then selected and matched against a database using the ARB Probe Match software. The ARB probe design tool has been used to design low-density custom-made POAs, composed of only a few hundreds of probes [23–25]. However, this probe design software is not well suited for large scale oligonucleotide probe design. Furthermore, it allows selecting only probes targeting known organisms and does not allow selecting explorative probes.

ARB and PRIMROSE tools allow selecting promising probes or primers for a single organism or a group of related organisms. However, emerging experimental approaches seek to simultaneously detect numerous organisms of interest thereby requiring the identification of large numbers of compatible probes [7, 26].

Oligonucleotide retrieving for molecular applications (ORMA) [27] is one of the most recent software proposed to select oligonucleotide probes. ORMA is composed of a set of scripts developed under Matlab and uses the BLAST program to check the specificity of the oligonucleotide probes selected. It allows designing probes for molecular application experiments on sets of highly similar sequences. ORMA was first applied to the design of probes targeting 16S rRNA genes, but it can also be used on any set of highly correlated sequences. This probe design tool has been used to design the HTF-Microbi-Array allowing high taxonomic level fingerprinting of the human intestinal microbial community [28].

All of these programs allow selecting probes targeting only known microbial communities with available sequences in public environmental databases. A few tools such as PhylArray [29] were designed with the possibility of selecting explorative probes for phylogenetic microarrays. PhylArray was developed with the Perl language. It allows selecting probes for a group of SSU rRNA sequences to globally survey known and unknown bacterial communities. Probe selection using PhylArray can take several days for only one large group of sequences.

In this work, we present a new parallel approach to select both known and explorative oligonucleotide probes on computing grids. The probe design strategy is based on the original algorithm PhylArray described in Militon et al. [29].

3. Material and Methods

3.1. Implementation. Our method was implemented in a program called PhylGrid 2.0. It was developed under Linux CentOS 5.4 with C++ and Perl. It uses three other programs: BLAST [6], Clustalw-MPI [30], and Opal [31].

Our approach hides the EGI grid to the user who just uses a regular computer which acts as a grid UI (User Interface: a grid component for user access to the grid). The first step was to implement the software on the User Interface (UI). This allows a direct connection to the EGI grid using a proxy authentication for the submission of multiple jobs. The main resources used by our grid application are the Workload Management System (WMS), a Berkeley Database Information Index (BDII), Computing Elements (CEs), and Storage Elements (SEs). We used the gLite middleware API commands. Submission, jobs management, and file transfer were implemented.

3.2. SSU rRNA Database Building. Probe design requires building a SSU rRNA database used as input and also as a reference database to check the specificity of all possible probes. This database must be of high quality in order to obtain the right design and to avoid wrong cross-hybridization results caused by poor sequences quality and erroneous affiliation in public environmental databases. Here, we developed a new algorithm to revisit, for more precision, the initial database described in Militon et al. [29].

All SSU sequences of the taxonomic divisions Prokaryotes (PRO), fungi (FUN), and environmental samples (ENV) downloaded from the European Molecular Biology Laboratory (EMBL) nucleotide sequence database were used as a reference to build our database carefully crafted for our probe design software. Several steps were needed. First, small subunit rRNA gene sequences (16S for prokaryotes and 18S for fungi) were extracted and filtered according to their quality and size. We kept only the sequences that met the following criteria.

(i) The sequence length is greater than 1,200 bases.

(ii) The sequence length is smaller than 1,600 bases for prokaryotic sequences and 1,800 bases for fungal sequences.

(iii) The sequence is assigned to the genus level in EMBL database (taxonomic information is extracted from the (OC) organism classification EMBL field).

(iv) The percentage of unknown nucleotides (not {A, C, G, T}) in the sequence is less than 1%.

(v) The maximum number of consecutive unknown bases must not exceed 5. The last two criteria allow removing low quality sequences.

These stringent parameters were chosen in order to allow an efficient probe design. Then, extracted sequences were grouped at the genus taxonomic rank and each group was included in its specific kingdom (prokaryote or fungi) according to the NCBI taxonomy database.

The next step consists in checking the orientation of the obtained sequences. We used BLASTN program and a reference sequence to check and correct the orientation of sequences that had been incorrectly oriented in the EMBL database.

Subsequently, a BlastClust was made on each group to eliminate redundant sequences, using the following parameters allowing a single-linkage clustering at 100% identity cutoff:

(i) -p F (nucleotide sequences);

(ii) -S 100 (similarity threshold);

(iii) -L 1 (minimum length coverage);

(iv) -b F (required coverage as specified by -L and -S on only one sequence of a pair).

Finally, for each group, we checked the homogeneity of its sequences. The aim was to eliminate sequences badly annotated and to create a homogeneous group of sequences to allow selecting specific probes for this group. This step was done using a modified version of Clustalw [32] to compute distances between sequences and the K-means method [33] to clustering sequences.

We used this algorithm to build a 16S rRNA database at the genus level. We obtained 2,069 prokaryotic genera; each is composed of a set of homogeneous sequences representing

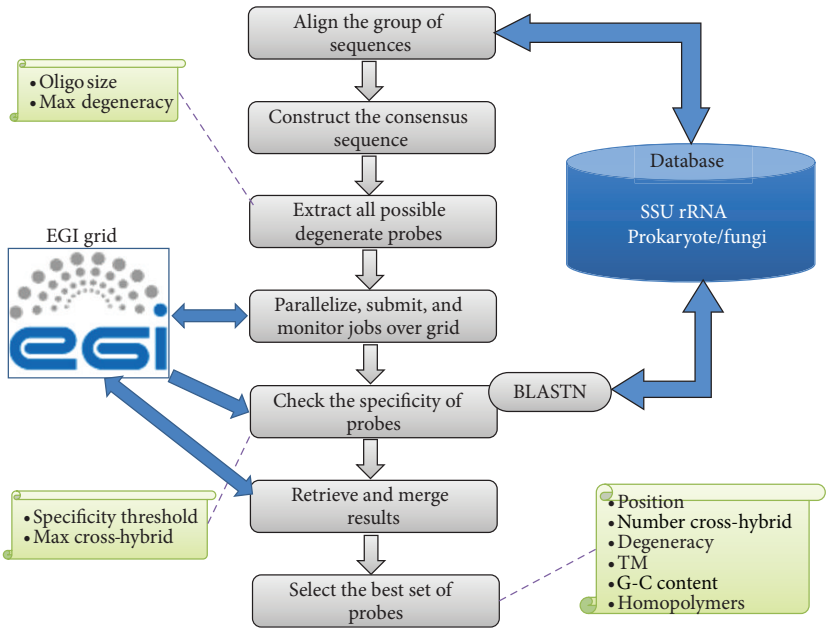


Figure 1: Summary of algorithm steps.

the whole diversity. Our algorithm can be easily adapted and used to build high-quality SSU rRNA databases for different taxonomic ranks (family, order, class, etc.).

3.3. The Probe Design Algorithm. Our algorithm uses 4 main input parameters: probe length, maximum degeneracy of a consensus probe, specificity threshold (the minimum value used to determine if the probe may hybridize with a nontarget sequence), and maximum number of cross-hybridizations. Figure 1 shows the different steps of our algorithm linked to the EGI grid.

To design probes for an input group of sequences selected by the user, a multiple sequence alignment is first made. For small groups of sequences, Clustalw-MPI [30] is used to align the sequences of the given group. However, for large groups of sequences, the multiple alignment is made in three steps to improve its quality and speed. First, BlastClust is made on each large group (using the parameters -L .98, -S 98, -p F, and -b F) to construct main subgroups of highly similar sequences. Then, sequences of each subgroup are aligned using Clustalw-MPI. Finally, Opal [31] is used to merge the obtained alignments and to reconstitute a complete alignment for the whole group.

The alignment file created is then used to construct a consensus sequence using the IUPAC degenerate nucleotide codes [34]. The aim is not only to obtain a common sequence that entirely represents the whole group of sequences targeted but also to improve alignment and correct possible sequencing errors. For example, in each column of the alignment representing a molecular site, if the number of unknown nucleotides ("N" or gap "-") is less than half the number of sequences aligned, all the unknown bases of the aligned sequences, at this position, are replaced by the specific or degenerate base calculated from all the specific nucleotides of this position. Else a gap "-" is inserted in the consensus sequence at this position.

The next step of the probe design strategy consists in incrementing a window of length "l" (l is the length of probes set by the user) along the consensus sequence to find all possible degenerate probes that do not contain gaps ("-") and whose degeneracy does not exceed the threshold value of maximum degeneracy allowed.

Subsequently, a parallelization is made to distribute all the extracted degenerate probes into "N" jobs (N is the number of jobs set by the user). For each job, all the degenerate probes are processed. Otherwise, all possible specific and explorative oligonucleotide probes are generated from each degenerate probe, using IUPAC codes [34]. These oligonucleotides are checked for cross-hybridizations against the reference SSU rRNA database initially built, using BLASTN program with the following parameters: "-W 7 -F F -S 1 -e 100 -b 20000".

Finally, all the obtained regular and explorative oligonucleotide probes are regrouped and saved in a final result file. For each degenerate or specific probe, all the associated information is provided, such as position, degeneracy, number and list of cross-hybridizations, and mismatch positions.

3.4. Parallelization Method. Selecting probes for a group of nucleic acid sequences and checking the specificity of each possible probe against a complete SSU rRNA database require a very important computation time. Our software allows running this kind of design on a computing grid. First, the user must choose the number of jobs to use. The consensus sequence, constructed from the alignment file

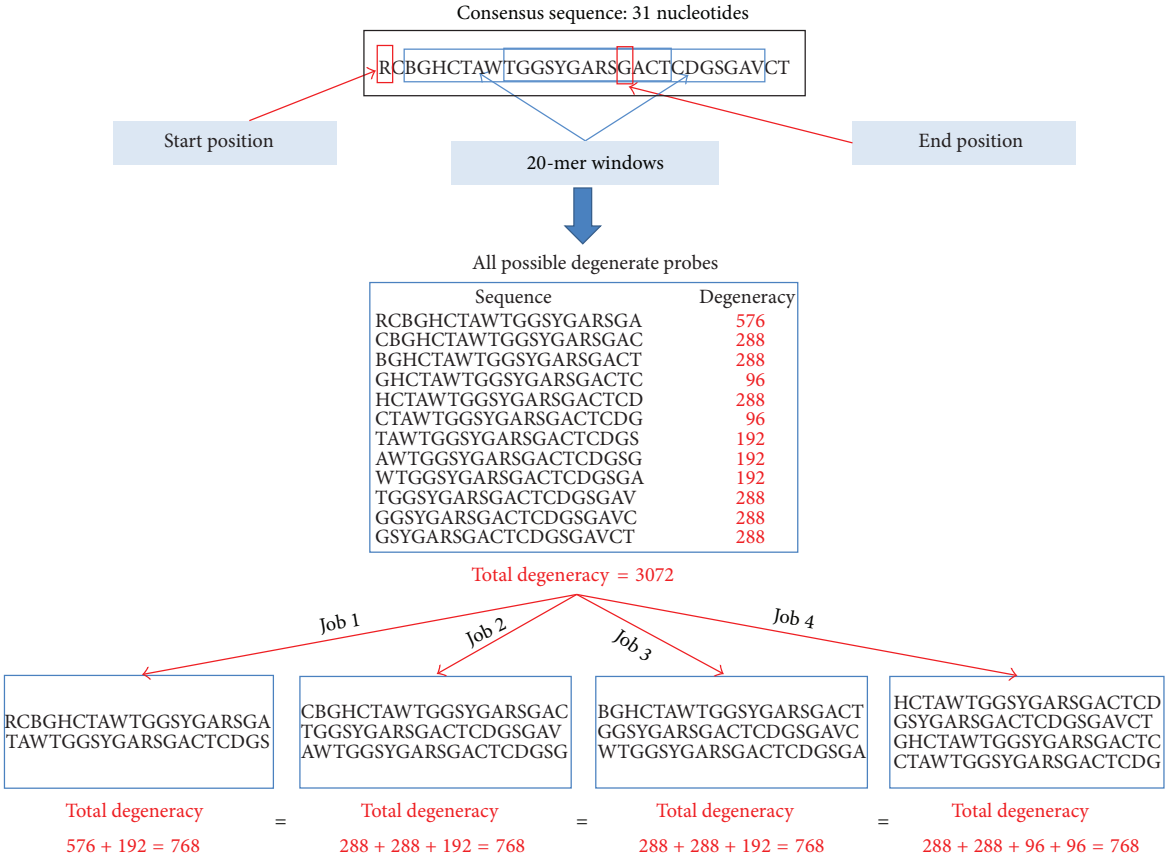


Figure 2: Parallelization strategy to define and submit jobs over the grid.

of each group of sequences, is read to extract all possible degenerate probes that do not contain gaps ("-"), based on the probe length set by the user. The degeneracy of each degenerate probe is calculated. If this degeneracy is less than "maximum degeneracy authorized by the user" (MaxDeg), the degenerate probe is saved. A weight value is calculated for each saved degenerate probe based on its degeneracy.

Once this step is performed, all valid degenerate probes saved are collected and put in the same file. This file must then be cut into "N" subfiles (N is the number of jobs set by the user) depending on the weight value of each degenerate probe and the sum of all the weight values. First, all the degenerate probes are sorted in descending order based on their weights. The mean degeneracy per subfile is then calculated based on the sum of all the weight values and the number of jobs desired. Finally, a "worst fit" algorithm [35] is used to put each degenerate probe in the largest possible free block in which this degenerate probe can be saved according to its weight. This method allows avoiding the creation of small unusable blocks by making the remainder as large as possible with the aim of making this remainder able to contain other degenerate probes. The subfiles created will have almost the same weight (Figure 2) and the same number of potential probes. Each subfile represents a job that will be submitted to the EGI computing grid.

Moreover, we have developed job monitoring scripts, with resubmission in case of failure to improve the reliability of our grid software. Three cases can be distinguished.

(i) The job submission failed: the job is resubmitted when a network route is found.

(ii) The job is submitted successfully and failed when executed: a new job is created and submitted.

(iii) The job is submitted successfully and done successfully but the other jobs are not finished: the program waits for all jobs and then merges all results in a single output file.

For running conditions, the database is copied on grid Storage Elements (SEs) accessible to all the grid jobs of a probe design. Regarding submission time, it is important not to overload the Workload Management System (WMS). Otherwise, the program may wait until each job is entirely associated with a CE of the EGI grid before submitting the next job. The following elementary configuration files are necessary to submit jobs successfully on the EGI grid.

Table 1: A comparison of the performance of the alignment method used in our software with that used in PhylArray [29], using 100 cores.

Aligned group	Number of sequences	Number of subgroups	Alignment time (seconds)		Speedup
			PhylArray	PhylGrid 2.0	
Vibrio	1,174	37	2,542	1,247	2.03
Bacillus	3,947	58	12,586	3,130	4.02

(i) JDL files: each job needs a job description language (JDL) file to be submitted on the Grid.

(ii) Script files: such files describe the elementary tasks that will be executed on the grid. The scripts contain operating system commands and Perl scripts called to perform probe design among all extracted degenerate probes. During execution, SSU rRNA database and subfiles containing degenerate probes are copied on the CE in which the job is running, and Blastn analysis is launched to test cross-hybridization.

Finally, the program is designed to be extensible by separating independently jobs in distinct designs. It creates a single data identifier for each probe design.

4. Results

In this section, we present the results obtained by our software on real data sets. We show the performance of our parallelization method compared to the original program PhylArray [29].

4.1. Database Building. We developed a new algorithm for the construction of a high-quality curated phylogenetic database, as described above. Our algorithm can be easily adapted and used to build high-quality SSU rRNA databases for different taxonomic ranks (genus, family, order, class, etc.). We used this algorithm to build a SSU rRNA database at the genus level. We obtained about 66,000 16S rRNA gene sequences representing 2,069 prokaryotic genera; each is composed of a set of homogeneous sequences representing the whole diversity. We used PhylGrid 2.0 and this database to create a complete phylogenetic oligonucleotide database composed of about 20,000 probes targeting 2,069 prokaryotic genera.

4.2. Alignment of Alignments. Dealing with the multiple sequence alignment for large groups of sequences, an alignment of alignments is achieved to improve the quality and speed of alignment. The alignment time is given in Table 1 for different groups with a varying number of sequences.

For instance, the performance of this method is 4 times faster than a simple multiple alignment when aligning the bacteria genus group "*Bacillus*."

4.3. Load Balancing Method. To distribute the probe design task on all used jobs equitably, we developed a load balancing method based on the degeneracy of all possible degenerate probes extracted from the consensus sequence constructed. To test the efficiency of our method, we compared it to the load balancing method used in the original algorithm PhylArray [29] that selects probes on a computing cluster. To distribute the computation on N processors, PhylArray splits the consensus sequence into N equal parts. Each part is then processed on a processor.

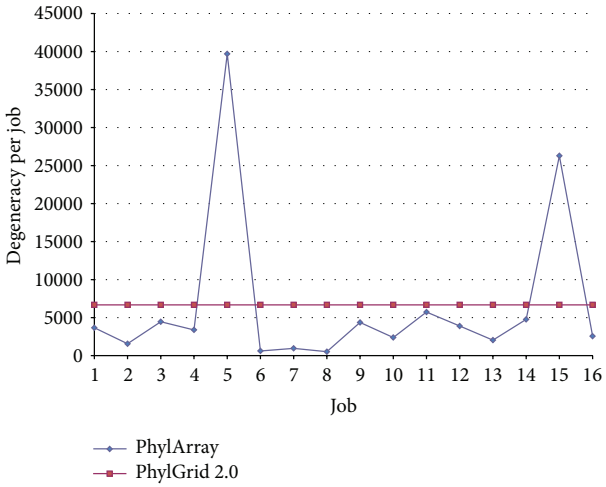


Figure 3: A comparison of our load balancing method with PhylArray [29] using 16 processors to select probes for "*Citrobacter*" group.

The comparison tests were made on real data sets, using respectively 16 jobs to select probes for the genus group "*Citrobacter*" (Figure 3), 8 jobs to select jobs for the genus group "*Haemophilus*" (Figure 4), and finally using 4 jobs to select jobs for 3 genus groups: "*Citrobacter*," "*Eubacterium,*" and "*Haemophilus*" (Table 2). This comparison shows that our method is more efficient than PhylArray. Using our method the different parts of the probe design, which processed each one on a processor, have almost the same value of degeneracy that is very close to the value of the mean degeneracy per job. For instance, as showed in Table 2, the load standard deviation between jobs is very small (0.5 probe) when using PhylGrid 2.0 compared to the high standard deviation obtained when using PhylArray (18,647.85 probes).

4.4. Use of the European Grid EGI. Our software allows users to submit parallel jobs to the EGI computing grid from Biomed Virtual Organization for the purpose of designing probes. To test the performance of our approach, we launched probes design for 10 prokaryotic genus groups simultaneously ("*Eubacterium*," "*Citrobacter*," "*Propionibacterium*," "*Neisseria*," "*Campylobacter*," "*Arcanobacterium*," "*Haemophilus*,"

Table 2: A Comparison of our load balancing method with PhylArray [29] using 4 processors to select probes for 3 genus groups.

Group	*Citrobacter*		*Eubacterium*		*Haemophilus*	
Software	PhylArray	PhylGrid 2.0	PhylArray	PhylGrid 2.0	PhylArray	PhylGrid 2.0
Mean degeneracy	26,722.75	26,722.75	37,132.25	37,132.25	20,100.75	20,100.75
Degeneracy job 1	13,068	26,723	41,435	37,133	28,600	20,101
Degeneracy job 2	41,782	26,723	43,466	37,132	32,335	20,101
Degeneracy job 3	16,381	26,723	10,273	37,132	4,314	20,101
Degeneracy job 4	35,660	26,722	53,355	37,132	15,154	20,100
Standard deviation	**14,142.836**	**0.5**	**18,647.85**	**0.5**	**12,853.09**	**0.5**

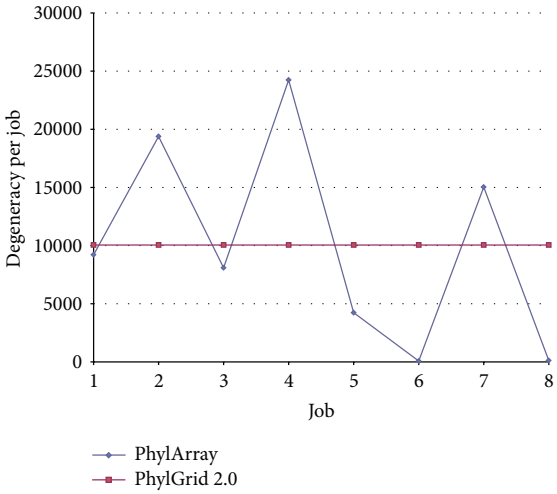


Figure 4: A comparison of our load balancing method with PhylArray [29] using 8 processors to select probes for "*Haemophilus*" group.

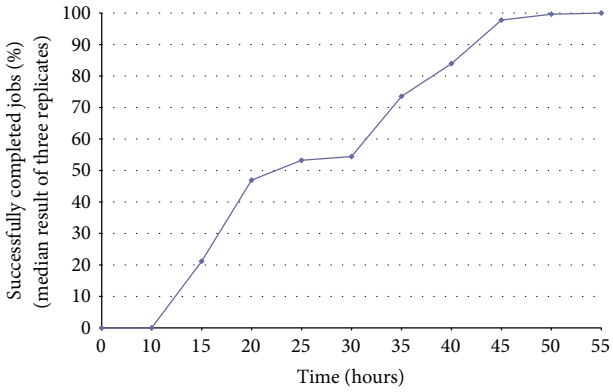


Figure 5: The median execution result of probe selection for 10 genus groups on the EGI grid using 586 jobs.

"*Kaistobacter*," "*Bacteriovorax*," and "*Riemerella*"), using the following parameters:

(i) probe length = 25;

(ii) specificity threshold = 0.88 (the probe must not have a similarity greater than or equal to 88%, with a nontargeted sequence);

(iii) maximum number of cross-hybridizations = 100;

(iv) maximum degeneracy = 2000.

This task needs more than 8 months to be processed on a single CPU core. We have launched probe designs for these groups on the EGI grid using a total of 586 jobs. We have repeated this test 3 times and the median result in terms of computational time was considered. Finally, we obtained all results successfully after less than 55 hours (with submission and waiting latency). Results are illustrated in Figure 5.

The obtained performance is here of about 106x for 586 jobs despite the submission and waiting latency of the EGI grid. Jobs submitted to a grid spend hours waiting in queues. The unavailability of some grid resources such as a Computing Element or a Storage Element can also cause the loss or blockage of jobs. This can of course increase the global computing time of our software which will however resubmit failed and lost jobs. For instance, in Figure 5, we can see a small decrease in throughput of returned completed jobs in the time window between 20 and 30 hours. This is due to the important resubmission of failed jobs at this computing phase. These jobs were submitted successfully at the beginning, but they failed or were blocked when executed.

5. Conclusions

In this work, we show that it is possible to select probes at large scale on a grid infrastructure with significant performance gains, without any particular grid submission optimizations (such as using pilot jobs). Our software allows selecting both specific and explorative (discovery of possible new species) probes with respect to excellent sensitivity and specificity. It takes advantage of the computing power offered by the EGI grid to propose at once probe design for thousands of groups. We also developed job monitoring scripts to improve the reliability and efficiency of our grid software.

The design of oligonucleotide probe on a computing grid requires optimizing the distribution of the probe design algorithm. This is why we developed an efficient parallelization method based on the degeneracy of all possible degenerate probes extracted from the consensus sequence that represents the input group. The probe selection is equally distributed

over a given number of jobs. We have compared our parallelization method with the original algorithm PhylArray [29]. We have shown that our approach is more efficient and allows a fine load balancing by sharing equitably the processing of probe selection for the input group across jobs. The comparison results of our load balancing method with that used in PhylArray—for a probe design with a mean degeneracy per job equal to 37,132.25 probes—showed that our software allowed creating jobs with a small load standard deviation of only 0.5 probe while PhylGrid generated a high load standard deviation of 18,647.85 probes between jobs. The experimental results obtained have shown that the parallel implementation of our software had significantly increased performance up to 106x when running around 600 jobs on the European Computing Grid (with submission and waiting latency). The performance of our software depends on the grid resource availability and also on the number and the size of designs that can be simultaneously launched. Hence, we have to consider Grid Computing only for large designs; otherwise, the queue waiting time and the time of data transfer on and to the grid can far exceed the computing time. For small groups of sequences, the use of a computing cluster or a multiprocessor will be more efficient than the use of a grid infrastructure for latency reasons. In our case, if we do not have tens of jobs with a job running time around 12 hours, we estimate that it is not worth submitting jobs to a computing grid where our jobs may queue for hours; instead our software suggests to consider local submissions to computing clusters.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank the Auvergne Regional Council and the European Regional Development Fund for the funding of Faouzi Jaziri scholarships. Nicolas Parisot was supported by the French "Direction Générale de l'Armement" (DGA). This work was also supported by the program Investissements d'avenir AMI 2011 VALTEX.

References

[1] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on Earth and in the ocean?" *PLoS Biology*, vol. 9, no. 8, Article ID e1001127, 2011.

[2] P. G. Falkowski, T. Fenchel, and E. F. Delong, "The microbial engines that drive Earth's biogeochemical cycles," *Science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[3] E. Dugat-Bony, E. Peyretaillade, N. Parisot et al., "Detecting unknown sequences with DNA microarrays: explorative probe design strategies," *Environmental Microbiology*, vol. 14, no. 2, pp. 356–371, 2012.

[4] J. Zhou and D. K. Thompson, "Challenges in applying microarrays to environmental studies," *Current Opinion in Biotechnology*, vol. 13, no. 3, pp. 204–207, 2002.

[5] M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore, "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays," *Nucleic Acids Research*, vol. 28, no. 22, pp. 4552–4557, 2000.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[7] D. Zhu, Y. Fofanov, R. C. Willson, and G. E. Fox, "A parallel computing algorithm for 16S rRNA probe design," *Journal of Parallel and Distributed Computing*, vol. 66, no. 12, pp. 1546–1551, 2006.

[8] E.-G. Talbi and A. Y. Zomaya, *Grid Computing For Bioinformatics and Computational Biology*, vol. 1 of *Wiley Series in Bioinformatics*, John Wiley & Sons, New York, NY, USA, 2007.

[9] I. Foster and C. Kesselman, *The Grid 2: Blueprint for a New Computing Infrastructure*, vol. 1, Morgan Kaufmann, Boston, Mass, USA, 2004.

[10] N. Jacq, C. Blanchet, C. Combet et al., "Grid as a bioinformatic tool," *Parallel Computing*, vol. 30, no. 9-10, pp. 1093–1107, 2004.

[11] M. Wagner, H. Smidt, A. Loy, and J. Zhou, "Unravelling microbial communities with DNA-microarrays: challenges and future directions," *Microbial Ecology*, vol. 53, no. 3, pp. 498–506, 2007.

[12] A. Loy, A. Lehner, N. Lee et al., "Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment," *Applied and Environmental Microbiology*, vol. 68, no. 10, pp. 5064–5081, 2002.

[13] E. L. Brodie, T. Z. DeSantis, D. C. Joyner et al., "Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation," *Applied and Environmental Microbiology*, vol. 72, no. 9, pp. 6288–6298, 2006.

[14] E. L. Brodie, T. Z. DeSantis, J. P. M. Parker, I. X. Zubietta, Y. M. Piceno, and G. L. Andersen, "Urban aerosols harbor diverse and dynamic bacterial populations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 1, pp. 299–304, 2007.

[15] T. C. Hazen, E. A. Dubinsky, T. Z. DeSantis et al., "Deep-sea oil plume enriches indigenous oil-degrading bacteria," *Science*, vol. 330, no. 6001, pp. 204–208, 2010.

[16] N. Weinert, Y. Piceno, G.-C. Ding et al., "PhyloChip hybridization uncovered an enormous bacterial diversity in the rhizosphere of different potato cultivars: many common and few cultivar-dependent taxa," *FEMS Microbiology Ecology*, vol. 75, no. 3, pp. 497–506, 2011.

[17] K. M. Handley, K. C. Wrighton, Y. M. Piceno et al., "High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment," *FEMS Microbiology Ecology*, vol. 81, no. 1, pp. 188–204, 2012.

[18] K. E. Ashelford, A. J. Weightman, and J. C. Fry, "PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database," *Nucleic Acids Research*, vol. 30, no. 15, pp. 3481–3489, 2002.

[19] S. Fraune, R. Augustin, F. Anton-Erxleben et al., "In an early branching metazoan, bacterial colonization of the embryo is controlled by maternal antimicrobial peptides," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 42, pp. 18067–18072, 2010.

[20] K. Bers, K. Sniegowski, P. Albers et al., "A molecular toolbox to estimate the number and diversity of Variovorax in the environment: application in soils treated with the phenylurea herbicide linuron," *FEMS Microbiology Ecology*, vol. 76, no. 1, pp. 14–25, 2011.

[21] D. Blaskovic and I. Barák, "Oligo-chip based detection of tick-borne bacteria," *FEMS Microbiology Letters*, vol. 243, no. 2, pp. 473–478, 2005.

[22] W. Ludwig, O. Strunk, R. Westram et al., "ARB: a software environment for sequence data," *Nucleic Acids Research*, vol. 32, no. 4, pp. 1363–1371, 2004.

[23] A. Loy, C. Schulz, S. Lücker et al., "16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order 'Rhodocyclales'," *Applied and Environmental Microbiology*, vol. 71, no. 3, pp. 1373–1386, 2005.

[24] H. Sanguin, A. Sarniguet, K. Gazengel, Y. Moënne-Loccoz, and G. L. Grundmann, "Rhizosphere bacterial communities associated with disease suppressiveness stages of take-all decline in wheat monoculture," *New Phytologist*, vol. 184, no. 3, pp. 694–707, 2009.

[25] M. R. Liles, O. Turkmen, B. F. Manske et al., "A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction," *Soil Biology and Biochemistry*, vol. 42, no. 5, pp. 739–747, 2010.

[26] D. Zhu, Y. Fofanov, R. C. Willson, and G. E. Fox, "ProkProbePicker (PPP): a fast program to extract 16S rRNA-targeted probes for prokaryotes," in *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '05)*, pp. 41–47, Las Vegas, Nev, USA, June 2005.

[27] M. Severgnini, P. Cremonesi, C. Consolandi, G. Caredda, G. De bellis, and B. Castiglioni, "ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design," *Nucleic Acids Research*, vol. 37, no. 16, p. e109, 2009.

[28] M. Candela, C. Consolandi, M. Severgnini et al., "High taxonomic level fingerprint of the human intestinal microbiota by Ligase Detection Reaction—Universal Array approach," *BMC Microbiology*, vol. 10, article 116, 2010.

[29] C. Militon, S. Rimour, M. Missaoui et al., "PhylArray: phylogenetic probe design algorithm for microarray," *Bioinformatics*, vol. 23, no. 19, pp. 2550–2557, 2007.

[30] K.-B. Li, "ClustalW-MPI: ClustalW analysis using distributed and parallel computing," *Bioinformatics*, vol. 19, no. 12, pp. 1585–1586, 2003.

[31] T. J. Wheeler and J. D. Kececioglu, "Multiple alignment by aligning alignments," *Bioinformatics*, vol. 23, no. 13, pp. i559–i568, 2007.

[32] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[33] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithms: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[34] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984," *Nucleic Acids Research*, vol. 13, no. 9, pp. 3021–3030, 1985.

[35] D. S. Johnson, "Fast algorithms for bin packing," *Journal of Computer and System Sciences*, vol. 8, no. 3, pp. 272–314, 1974.

1.4 Discussion

Malgré le nombre important de logiciels de sélection de sondes oligonucléotidiques pour biopuces ADN (Lemoine *et al.* 2009 ; Dugat-Bony *et al.* 2012b ; Parisot *et al.* 2014), beaucoup ne sont pas adaptés pour répondre à des problématiques environnementales. La construction de biopuces phylogénétiques environnementales nécessite de travailler avec des masses de données génomiques telles qu'il est nécessaire de proposer de nouvelles stratégies innovantes. PhylGrid 2.0, en s'appuyant sur la grille de calculs européenne, représente le premier logiciel de détermination de sondes à grande échelle. Grâce à cette stratégie, il est possible de filtrer rapidement des millions de sondes candidates pour ne conserver que les sondes de haute qualité.

Travailler sur une architecture à grande échelle comme la grille de calculs implique néanmoins une réflexion particulière sur la gestion des ressources. En effet, de nombreux *jobs* envoyés sur la grille de calculs n'aboutissent pas, certains peuvent i) être bloqués en raison de l'absence de ressources disponibles, ii) échouer en cours d'exécution à cause de pannes diverses ou iii) être perdus après la soumission sans jamais être exécutés. Certains auteurs ont observés jusqu'à 33% d'échecs lors de la soumission de *jobs* à une grille de calculs (Li *et al.* 2006). Les algorithmes développés doivent donc être capables d'identifier les *jobs* ayant échoué pour pouvoir les resoumettre rapidement. Certaines stratégies visent même à soumettre plusieurs fois un même *job* pour s'assurer d'obtenir le résultat du traitement.

La grille de calculs étant également une structure partagée par de nombreux scientifiques, ses ressources ne sont donc pas toujours disponibles et les temps d'attentes sont fluctuants. Tous ces paramètres doivent être pris en compte avant de choisir d'utiliser ce type d'architecture. Ainsi, pour une détermination de sondes ponctuelle, dirigée vers quelques genres bactériens par exemple, il est préférable d'utiliser d'autres architectures comme les *clusters* de calculs. On estime que l'utilisation de la grille de calculs est rentable à partir d'une dizaine de *jobs* d'au moins 12 heures chacun.

Il est donc nécessaire de poursuivre les développements bioinformatiques dans le domaine de la conception de sondes afin de proposer des stratégies innovantes applicables à moindre échelle aux problématiques environnementales.

2. Développement d'un logiciel de sélection de sondes oligonucléotidiques : KASpOD

2.1 Contexte

Du fait de l'augmentation constante du nombre de séquences déposées dans les bases de données qui doivent être prises en compte, les logiciels de détermination de sondes doivent être de plus en plus performants. A l'heure actuelle, la plupart de ces logiciels ne permettent pas de gérer ces grands jeux de données. PhylGrid 2.0 (cf. §1. Amélioration et déploiement sur la grille de calculs d'un logiciel de détermination de sondes oligonucléotidiques pour biopuces phylogénétiques : PhylGrid 2.0) a permis de réduire les problèmes engendrés par cette masse de données génomiques en s'appuyant sur une architecture de calculs à grande échelle. Néanmoins, une des étapes clés de la détermination de sondes oligonucléotidiques spécifiques d'un groupe ciblé regroupant un grand nombre de séquences reste délicate et ce peu importe la puissance de calculs.

En effet, l'étape d'alignement multiple des séquences ciblées, permettant d'identifier des régions conservées dont seront extraites les sondes, devient rapidement irréalisable pour des jeux de données de plusieurs milliers de séquences. De nouvelles stratégies basées sur l'utilisation des *k*-mers ont donc été développées et permettent de contourner cette contrainte (Bader *et al.* 2011 ; Hysom *et al.* 2012).

2.2 Objectif

Une nouvelle stratégie pour la détermination de sondes dédiées à l'écologie microbienne a donc été envisagée. L'objectif était de proposer un nouveau logiciel offrant la possibilité d'effectuer la détermination des sondes à partir d'une masse importante de données et donc de s'affranchir de l'étape d'alignement multiple. De plus, ce logiciel devait permettre de sélectionner des sondes oligonucléotidiques de qualité tout en intégrant le caractère exploratoire.

La stratégie mise en place repose sur i) l'identification de mots (*k*-mers) longs (entre 18 et 31-mers) retrouvés, de manière exacte, uniquement dans le groupe de séquences ciblées (et absents d'un groupe non ciblé également fourni par l'utilisateur) ; ii) le regroupement des *k*-mers d'un même groupe afin d'aligner les mots provenant strictement de la même région génomique. Ainsi, à partir de l'alignement multiple de chaque groupe, un mot consensus peut

être défini en intégrant la variabilité génomique à chaque site moléculaire. Cet oligonucléotide consensus peut donc être dégénéré à certaines positions rendant ainsi possible la détermination de signatures exploratoires. Contrairement à PhylArray ou Phylgrid 2.0 qui imposent de tester chaque combinaison non dégénérée, la stratégie mise au point utilise directement cet oligonucléotide dégénéré pour le test de spécificité mais également pour évaluer la couverture de la sonde. La couverture d'une sonde désigne sa capacité à s'hybrider avec les séquences ciblées, et permet une évaluation *in silico* de la sensibilité. Pour cela, la spécificité est évaluée par comparaison avec le groupe de séquences non ciblées alors que la couverture utilise le groupe de séquences ciblées. Il est possible de fixer un nombre maximal de différences autorisées entre la signature et la séquence cible ou non-cible. Ainsi, si cette distance est fixée à deux par l'utilisateur, l'ensemble des séquences du groupe cible présentant au maximum deux différences (mésappariements ou *gaps*) avec la signature testée est pris en compte pour le calcul de la couverture. De même, les séquences du groupe non-cible ayant jusqu'à deux différences sont considérées comme pouvant induire de potentielles hybridations croisées.

L'ensemble de la stratégie mise en place a conduit au développement du logiciel KASpOD pour « *K-mer based Algorithm for high-Specific Oligonucleotide Design* ».

2.3 Principaux résultats

Le travail réalisé a permis de proposer un nouveau logiciel de détermination de sondes oligonucléotidiques utilisant une stratégie originale basée sur les *k*-mers qui a donné lieu à une publication sous la forme d'une « *Applications Note* » dans le journal « *Bioinformatics* ». Le logiciel KASpOD est utilisable *via* une interface web (http://g2im.u-clermont1.fr/kaspod/). Pour le traitement de grandes masses de données, une version téléchargeable est également disponible sur ce même site pour une utilisation en local.

Afin d'évaluer la performance de ce nouvel outil, différentes sondes utilisées pour la construction d'une biopuce phylogénétique ont été déterminées, en ciblant l'ensemble des genres procaryotes présents au sein de la base de données Greengenes (McDonald *et al.* 2012). Cette base de données a été filtrée pour ne conserver que 252 183 séquences de haute qualité définissant 1295 genres procaryotes. Au total, 22 613 sondes non chevauchantes de 25-mers couvrant l'ensemble des séquences des genres ciblés ont été déterminées. Le temps de calculs nécessaire au logiciel pour la détermination des sondes est relativement court puisque les calculs soumis *via* l'interface web de KASpOD sont exécutés sur une machine

multi-processeurs gérée par le CRRI (Centre Régional de Ressources Informatiques). Grâce aux 135 processeurs de 2,2 GHz et 2 Go de RAM chacun, il faut compter, pour des sondes de 25-mers, et aucune différence autorisée (pour la couverture et la spécificité), environ 9 minutes pour un genre procaryote avec un nombre restreint de séquences (685), 36 minutes avec un jeu moyen (4733) et 53 minutes avec un jeu de séquences plus conséquent (9528). Toutefois ce temps de calculs est dépendant de certains paramètres comme le nombre maximal de différences autorisées entre la signature et le groupe cible ou non-cible. Ainsi, si cette distance est fixée à deux, les temps de calculs précédents sont respectivement de 55 minutes, 4,5 heures et 16 heures. D'autres paramètres influencent aussi le temps nécessaire à la détermination de sondes comme la diversité de séquences au sein du groupe ciblé. Une diversité importante implique un nombre de *k*-mers élevé et par conséquent des temps de calculs plus conséquents. Par ailleurs, la disponibilité du *cluster* de calcul est également un paramètre important. En effet, d'autres utilisateurs ont accès à cette machine et le temps passé en file d'attente peut être fluctuant.

D'une manière générale, KASpOD permet la sélection rapide de plusieurs sondes hautement spécifiques pour chaque groupe de séquences donné en entrée. L'utilisateur ayant plusieurs sondes à sa disposition, il a la possibilité de choisir la ou les meilleures sondes sur des critères thermodynamiques (T_m, %GC…), la position sur le gène ou encore la confiance accordée aux résultats obtenus (hybridations croisées).

Article n°4

KASpOD--a web service for highly specific and explorative oligonucleotide design.

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 23 2012, pages 3161–3162
doi:10.1093/bioinformatics/bts597

Data and text mining

Advance Access publication October 9, 2012

KASpOD—a web service for highly specific and explorative oligonucleotide design

Nicolas Parisot[1,2], Jérémie Denonfoux[1,2], Eric Dugat-Bony[1,3], Pierre Peyret[1,3] and Eric Peyretaillade[1,3,*]

[1]Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448 and [2]UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France and [3]Clermont Université, Université d'Auvergne, UFR Pharmacie, 63000 Clermont-Ferrand, France

Associate Editor: David Posada

ABSTRACT

Summary: KASpOD is a web service dedicated to the design of signature sequences using a *k*-mer–based algorithm. Such highly specific and explorative oligonucleotides are then suitable for various goals, including Phylogenetic Oligonucleotide Arrays.

Availability: http://g2im.u-clermont1.fr/kaspod.

Contact: eric.peyretaillade@udamail.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 13, 2012; revised on September 25, 2012; accepted on September 28, 2012

1 INTRODUCTION

Environmental DNA microarrays, including Phylogenetic Oligonucleotide Arrays (POAs), are key technologies that are well adapted to profiling environmental communities (Dugat-Bony *et al.*, 2012b). The extreme diversity of microorganisms, however, means that molecular community exploration or specific analysis of microbial groups are faced with a new challenge: designing group-specific probe sets that must harbour a high coverage (i.e. being able to hybridize with all the target sequences) and a high specificity, showing no cross-hybridizations with non-target sequences (Loy *et al.*, 2008). Sensitivity (i.e. being able to detect even low abundance targets) and uniformity (i.e. uniform thermodynamic behaviours for all the probes) are also main criteria in the selection of the best probe set (Wagner *et al.*, 2007).

The development of comprehensive POAs requires integrating large datasets produced by metagenomics projects to assess the coverage and specificity of the probe set. Unfortunately, many available probe design programmes are not suitable to deal with such data (Dugat-Bony *et al.*, 2012b). To overcome this limitation, two recent strategies have been implemented (Bader *et al.*, 2011; Hysom *et al.*, 2012). Despite major speed improvements, both strategies are still not able to define explorative probes. They only define regular oligonucleotides found uniquely in the target group, whereas explorative probes take into account the sequence variability within the target group to define new combinations not yet deposited in public databases but potentially present in the environment.

In spite of large amounts of data, our current vision of the microbial diversity is, indeed, still incomplete. This is partially explained by the tremendous diversity of microbial species, ecological niches and technological limits: detecting 90% of the richness in some complex environments could require tens of thousands of times the current sequencing effort (Quince *et al.*, 2008). Microarrays coupled with explorative probe design strategies are, therefore, well suited to survey complete microbial communities, including microorganisms with uncharacterized sequences (Dugat-Bony *et al.*, 2012a; Terrat *et al.*, 2010).

Currently, the only software dedicated to POAs that allows the design of explorative probes is the PhylArray programme (Militon *et al.*, 2007), which relies on group-specific alignments before the probe design step to identify conserved probe-length regions. Building large multiple sequence alignments, however, represents a time-consuming task that is not compatible with high-throughput data.

Here we propose KASpOD, a fast and alignment-free algorithm to detect group-covering signature sequences allowing the design of explorative probes.

2 METHODS

2.1 Usage

KASpOD takes as input a target sequence set and a database of non-target sequences. The web interface accepts two parameters to design signatures: the oligonucleotide length (18–31-mer), and the edit distance between signatures and full-length sequences to perform specificity and coverage evaluation steps. The edit distance is defined as the total number of differences, gaps and/or mismatches allowed between the probe and its target.

2.2 Algorithm

KASpOD consists of three computational stages (Fig. 1).

2.2.1 Search for group-specific k*-mers* The first stage is the extraction of every *k*-mer from both the target and the non-target groups by using Jellyfish version 1.1.4 (Marcais and Kingsford, 2011). For large target groups (>100 sequences), a noise-reduction step is performed to remove *k*-mers occurring only once. Every *k*-mer found in both groups is then removed from the signature candidates, as it occurs exactly in the non-target group.

*To whom correspondence should be addressed.

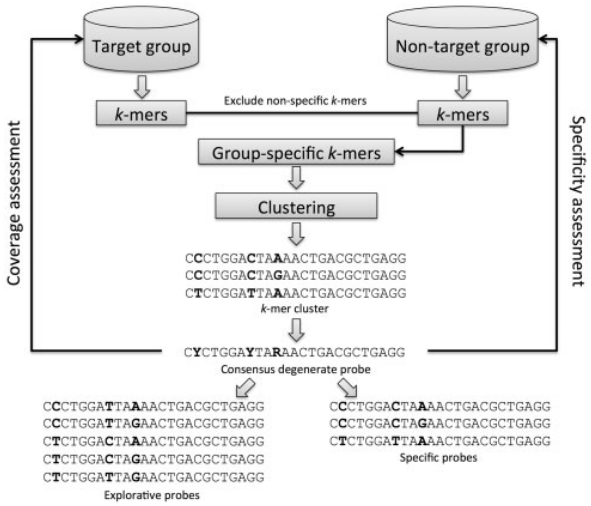


Fig. 1. The KASpOD programme workflow

2.2.2 Consensus signature sequences building The second stage consists of clustering fully overlapping *k*-mers using CD-HIT version 4.5.4 (Li and Godzik, 2006) at an 88% identity clustering threshold. For each cluster, a degenerate consensus signature is built taking into account sequence variability at each position.

2.2.3 Coverage and specificity evaluation The last stage performs a coverage assessment of each degenerate consensus *k*-mer against the target group, by using PatMaN version 1.2.2 (Prüfer *et al.*, 2008). Coverage is computed using the number of exact or non-exact (with at most the edit distance) matches in the target group. Specificity is assessed in the same way by comparing degenerate probes against the non-target group sequences.

3 RESULTS

We used KASpOD to design 25-mer probes for 1295 prokaryotic genera based on the recently published Greengenes taxonomy (McDonald *et al.*, 2012) (see Supplementary Data 1 for complete procedure). Finally, 22 613 group-specific signatures were designed (Supplementary Table 2) and are freely available on the KASpOD website (http://g2im.u-clermont1.fr/kaspod/about.php). This high-quality probe set could be used to build a POA to allow monitoring of complete prokaryotic communities in complex environmental samples. The probe set was not filtered using thermodynamic calculations, to let the users select the entire probe set, or subset, for their own applications, such as Polymerase Chain Reaction (PCR), Fluorescence In Situ Hybridization (FISH), gene capture or *in silico* for rapid sequence identification.

A runtime performance analysis of the web service has been performed and results are available in the Supplementary Data 3.

As KASpOD does not allow the generation of probes longer than 31 nucleotides, an interesting strategy would be to combine KASpOD and GoArrays (Rimour *et al.*, 2005) to concatenate two short probes with a random linker. This approach produces oligonucleotide probes as specific as short probes and as sensitive as long ones. KASpOD could, therefore, be used for applications such as Functional Genes Arrays, offering the opportunity to generate group-specific and explorative probes, allowing a broad coverage of multiple sequence variants for a given gene family.

ACKNOWLEDGEMENTS

The authors thank S. Terrat and A. Mahul for their help.

Funding: This work was supported by Direction Générale de l'Armement (DGA).

Conflict of Interest: none declared.

REFERENCES

Bader,K.C. *et al.* (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, **27**, 1546–1554.

Dugat-Bony,E. *et al.* (2012a) In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. *Microb. Biotechnol.*, **5**, 642–653.

Dugat-Bony,E. *et al.* (2012b) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environ. Microbiol.*, **14**, 356–371.

Hysom,D.A. *et al.* (2012) Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. *PLoS One*, **7**, e34560.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Loy,A. *et al.* (2008) probeCheck—a central resource for evaluating oligonucleotide probe coverage and specificity. *Environ. Microbiol.*, **10**, 2894–2898.

Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

McDonald,D. *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.

Militon,C. *et al.* (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics*, **23**, 2550–2557.

Prüfer,K. *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.

Quince,C. *et al.* (2008) The rational exploration of microbial diversity. *ISME J.*, **2**, 997–1006.

Rimour,S. *et al.* (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, **21**, 1094–1103.

Terrat,S. *et al.* (2010) Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics*, **11**, 478.

Wagner,M. *et al.* (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microb. Ecol.*, **53**, 498–506.

Supplementary Data 1: Technical details about the prokaryotic oligonucleotide array (POA) construction from input data management to the probe design.

Step 1: 16S rDNA database construction

The current release of Greengenes (09-May-2011) containing 406,997 sequences was downloaded and extracted from the following URL: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz.

Then, using a PERL script, only the sequences assigned to a genus were retained for further analyses. These 310,575 sequences were then sorted by genus into different FASTA files. For each genus, a clustering step was performed at a 100% identity threshold using CD-HIT in order to remove any redundancies. Moreover, only high-quality sequences were retained:

- Sequence length greater than 1,200 nucleotides
- Less than 1% of ambiguous nucleotides (N's)

After this processing pipeline, the 16S rDNA database contained 252,250 high-quality sequences. The clustering of the whole database at high-identity thresholds (99%, 98% and 97%) coupled with manual curation, allowed us to remove potentially badly assigned sequences. Furthermore, some microbial genera were clustered together, as they were hardly distinguishable on the basis of their sequences.

Eventually, 252,183 16S rDNA sequences were fed to KASpOD to perform the probe design.

Step 2: Probe design

Each genus was then used to perform a probe design with a stand-alone version of the KASpOD software. The non-target group was composed of the 252,183 16S rDNA sequences minus the target group (*i.e.* the genus being processed).

The smallest target groups contained only one sequence (*Arhodomonas*, *Methylosphaera*, *Roseisalinus*, *Subtercola* and *Thermopallium*) with a file size of 2KB, whereas the largest was the *Corynebacterium* genus with 20,093 sequences and a file size of 33MB.

Concerning the non-target groups, the largest was composed of 252,182 sequences with a file size of 401MB and the smallest contained 232,090 sequences and had a file size of 368MB.

Each genus represents one job and computations were distributed on a multi-processor computer (40CPUs). The whole design for the 1,295 microbial genera lasted 10 days.

Step 3: Probe selection

The last stage consists of the probe set selection from the 3,242,105 candidate probes previously generated. Using a PERL script, probes were selected in order to build a probe set where each of the selected 252,183 16S rDNA sequences were covered by at least three different probes.

First step: the non-overlapping probes are selected within the probes showing no cross-hybridisations.

Second step: while there are some 16S rDNA sequences which are not covered by at least three probes, the programme selects additional probes with increasing numbers of cross-hybridisations. During this step, the programme ensures that no more than two probes show significant cross-hybridisation with the same non-targeted genus, thereby avoiding misleading interpretations of hybridisation data.

Finally, 22,613 probes were selected which could be used to build a phylogenetic oligonucleotide microarray, or for other applications (PCR, qPCR, FISH, gene capture, in silico sequence identification).

Supplementary Table 3: Runtime performance analysis of the KASpOD's web-service.

Oligo length	Edit distance	Target File	Non-Target File	Time (minutes)
18	0	Small		6
		Medium	Large	32
		Large		53
			Small	39
		Large	Medium	69
			Large	53
	2	Small		78
		Medium	Large	360
		Large		778
			Small	733
		Large	Medium	832
			Large	778
25	0	Small		9
		Medium	Large	36
		Large		53
			Small	53
		Large	Medium	52
			Large	53
	2	Small		55
		Medium	Large	271
		Large		958
			Small	549
		Large	Medium	883
			Large	958
31	0	Small		9
		Medium	Large	18
		Large		53
			Small	52
		Large	Medium	54
			Large	53
	2	Small		25
		Medium	Large	201
		Large		789
			Small	575
		Large	Medium	656
			Large	789

Target Files: The small target file was composed of 685 16S rDNA sequences (1MB) belonging to the *Stenotrophomonas* genus. The medium target file was composed of 4,733 16S rDNA sequences (7.4MB) belonging to the *Faecalibacterium* genus. The large target file was composed of 9,528 16S rDNA sequences (15MB) belonging to the *Pseudomonas* genus.

Non-Target Files: The non-target files were built using a reduced personal 16S rDNA sequences database without the *Stenotrophomonas*, *Faecalibacterium* and *Pseudomonas* genera. The small non-target file was constructed by randomly taking 500 sequences out of the database (740KB). The medium non-target file was constructed by randomly taking 5,000 sequences out of the database (7.2MB). The large non-target file was constructed by randomly taking 10,000 sequences out of the database (14MB).

Nevertheless, the authors would like to emphasize that the run times are given for guidance and are dependent on many parameters (*e.g.* number of jobs on the cluster queue, heterogeneity of the target file or number of cross-hybridizations). The job status is therefore important for the user to know whether or not the job is running.

Ce travail a également fait l'objet d'un chapitre d'un ouvrage prochainement disponible et intitulé « *Microarray Technology and its Applications* », a été réalisé sous la direction du Dr. Lin Wang et du Dr. Paul C.H. Li de l'Université Simon Fraser (Canada).

Chapitre d'ouvrage n°2

Probe design strategies for oligonucleotide microarrays.

In, Microarray Technology and its Applications

Probe design strategies for oligonucleotide microarrays

Nicolas Parisot[1], Eric Peyretaillade[1], Eric Dugat-Bony[2], Jérémie Denonfoux[3], Antoine Mahul[4] and Pierre Peyret[1,*].

[1] *Clermont Université, Université d'Auvergne, EA 4678, CIDAM, BP 10448, F-63000 Clermont-Ferrand, France.*
[2] *INRA, AgroParisTech, UMR 782 Génie et Microbiologie des Procédés Alimentaires, Centre de Biotechnologies Agro-Industrielles, Thiverval-Grignon, France.*
[3] *Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, Lille, France.*
[4] *Clermont Université, CRRI, F63177 Aubière, France.*

[*] To whom correspondence should be addressed:

Pierre Peyret, EA 4678 CIDAM, 28 place Henri Dunant, F-63001 Clermont-Ferrand, France; E-mail: pierre.peyret@udamail.fr; Tel: +33 473 178 308; Fax: +33 473 275 624

Running head: Explorative probe design strategies

Abstract

Oligonucleotide microarrays have been widely used for gene detection and/or quantification of gene expression in various samples ranging from a single organism to a complex microbial assemblage. The success of a microarray experiment, however, strongly relies on the quality of designed probes. Consequently, probe design is of critical importance and should therefore consider multiple parameters in order to ensure high specificity, sensitivity, and uniformity as well as potentially quantitative power for each probe. Moreover, to assess the complete gene repertoire of complex biological samples as such those studied in the field of microbial ecology, exploratory probe design strategies must be also implemented to target not-yet-described sequences. To design such probes, two algorithms, KASpOD and HiSpOD, have been developed. Here we describe the use of these software via two user-friendly web services for designing oligonucleotide probes taking into account all the crucial parameters necessary for the design of highly effective probes especially in the context of microbial oligonucleotide microarrays.

Key words: KASpOD, HiSpOD, DNA microarrays, probe design, explorative probe

1. Introduction

The advancement of microarray technology (*e.g. in situ* synthesis technologies), coupled to the evolution of microarrays slide formats, led oligonucleotide microarrays to become high-throughput molecular tools. Holding millions of probes spotted on a single glass slide, the high-density oligonucleotide microarrays can monitor the presence and/or the expression, of thousands of genes, combining qualitative and quantitative aspects in only one experiment ***(1-3)***.

The success of a microarray experiment however strongly depends on the determination of the best probe set while taking the biological question into account. For instance, transcriptome arrays or whole-genome arrays (WGAs) target a single organism whose genome is sequenced whereas environmental DNA microarrays aim to study complex microbial mixtures of known and unknown microorganisms. Probe design is thus one of the most critical step because the selected oligonucleotide probe set will have to combine: i) sensitivity (*i.e.* probes should detect low abundance targets in complex mixtures), ii) specificity (*i.e.* probes should not cross-hybridize with non-target sequences) and iii) uniformity (*i.e.* probes should display similar hybridization behaviour) ***(4, 5)***.

Qualitative and quantitative improvements in the next-generation sequencing methods have produced a substantial volume of sequence information and will continue to accumulate large amounts of sequence data sets in public databases. It is now possible to take advantage of such sequencing data to develop comprehensive microarrays by using explorative probe design strategies. Such strategies offer the opportunity to survey both known and unknown sequences ***(6)***. Explorative probes strategy uses the sequence variability within the targeted sequences to define new combinations potentially present in natural environments and that have not yet been described and deposited in public databases.

Several software solutions are available to help the user and solve the current bottlenecks in the choice of high-quality probe sets (**Table 1**). Each software program has its own advantages and drawbacks, and the choice of programs must be made in total accordance with the nature of projects and the basic scientific question.

More recently microarrays were adapted in a flexible and easy-to-use manner to profiling environmental communities in the area of microbial ecology ***(5, 7)***. Indeed, designing oligonucleotide microarrays that can be used to survey the extreme diversity of microorganisms living in various ecosystems represents a stimulating challenge in the field of microbial ecology. Although several Whole-genome arrays (WGAs) have been developed in the last few years, Phylogenetic oligonucleotide arrays (POAs), targeting the SSU rRNA

genes, as well as Functional gene arrays (FGAs), targeting key genes encoding enzymes involved in metabolic processes, are the two major approaches to assess diversity of microbial communities in the environment ***(5)***.

1.1 Phylogenetic Oligonucleotide Arrays (POAs)

To rapidly characterise the members of microbial communities present in complex environments, numerous phylogenetic oligonucleotide arrays (POAs) have been developed targeting the SSU rRNA biomarker ***(8-15)***. Fully automated software and manual approaches have both been developed to design POAs (**Table 1**).

Such probe design strategies are generally based on aligned input data or on performing a multiple sequence alignment as the first step of the algorithm. To design probes with an optimal coverage of the target group, multiple sequence alignments are usually converted into consensus sequences that account for the sequence variability at each position. Then, probe design programs search for conserved regions to select the best oligonucleotides.

The first software program dedicated to POAs that offered the possibility of designing explorative probes was the PhylArray program ***(16)***. PhylArray was developed to survey whole microbial communities, including both known and unknown microorganisms, in complex environments. A degenerate consensus sequence is deduced from a multiple alignment of targeted SSU rRNA sequences. Degenerate candidate probes are then selected along the consensus sequence, and all the non-degenerate combinations deduced from the consensus are checked for cross-hybridisations against a SSU rRNA database. Among the combinations derived from each degenerate probe, some correspond to sequences that have not yet been deposited in public databases, namely explorative probes. Such probes should, therefore, allow the detection of undescribed microorganisms belonging to the targeted taxon.

Even if PhylArray was designed to account for all of the sequence variability within the targeted sequences, it is limited in its ability to manage large input datasets due to the fact that it relies on initials multiple sequence alignments. Consequently, new probe design strategies are needed to define explorative probes based on large databases.

1.1.1 Introducing KASpOD

The KASpOD ***(17)*** software, using a *k*-mer based strategy was developed to overcome this limitation. KASpOD (K-mer based Algorithm for highly Specific and explorative

Oligonucleotide Design) consists of three computational stages (**Fig. 1**). The user first provides two datasets that correspond to the targeted group of sequences and the non-target group. KASpOD will subsequently search probes that cover the target group minimizing the cross-hybridisation with the non-targeted sequences. The first stage of this algorithm is the extraction of every *k*-mer from the target and the non-target groups. Every *k*-mer found in both the target and the non-target groups is removed from the list of probe candidates. A clustering step is then performed to gather probes from the same genomic location. For each cluster, a degenerate consensus probe is deduced that accounts for the sequence variability within the cluster. Among the sequence combinations derived from each degenerate oligonucleotide, some correspond to sequences not previously included in the target group and therefore represent explorative probes. Finally, the last stage of the KASpOD algorithm consists of assessing the coverage (*i.e.* percentage of sequences within the targeted group matching with the oligonucleotide probe) and specificity of each degenerate consensus *k*-mer against the target and non-target groups.
KASpOD is provided as both a web service (http://g2im.u-clermont1.fr/kaspod/) and a stand-alone package. KASpOD is also not restricted to construct POAs and could be used to design probes for other microarrays (WGAs or FGAs)

1.1.2 Probe design parameters using KASpOD

KASpOD defines group-specific signatures based on two FASTA files: one containing the target group and the second with the non-target sequences. The user can choose the oligonucleotide length (between 18 and 31 nucleotides) as well as the edit distance to perform both coverage and specificity assessments. The edit distance is defined as the upper limit of tolerated differences (gaps and/or mismatches) between the probe and its target (or non-target). For a classical POA design we suggest 25mers probes and an edit distance threshold set to 2.

1.1.3 Probe design results using KASpOD

Once the probe design is done, KASpOD provides a downloadable CSV file containing the results. Each line represents a candidate probe and the columns correspond to: probe sequence, number of sequences in the target group, start, end, coverage (%), number of sequences in the non-target group and the number of cross-hybridisations. Start and end

positions in the results file are given for guidance since they are only defined according to the probe's best match in the target group.

1.2 Functional Gene Arrays (FGAs)

High-density oligonucleotide functional gene arrays (FGAs) provide the best high-throughput molecular tools to access the tremendous functional diversity of ecosystems ***(18)***. Although most strategies are limited to the determination of probes that target specific gene sequences within a single genome dataset, few strategies offer the opportunity to design probes that permit broad coverage of multiple sequence variants for a given gene family ***(6, 19)*** (**Table 1**).

1.2.1 Introducing HiSpOD

The HiSpOD (High Specific Oligo Design) program was developed ***(3)*** in this context of microbial ecology. HiSpOD (**Fig. 2**) allows designing both gene-specific and sequence-specific probes. Gene-specific probes are computed using consensus sequences obtained after multiple alignments of nucleic sequences belonging to the same gene family. All combinations deduced from the degenerate probes are then divided into regular probes for sequences available in databases, and explorative probes that represent putative new genetic signatures that do not correspond to any previously described sequence. Sequence-specific probes can also be designed through HiSpOD by using non-degenerate nucleic acid sequences corresponding to a unique gene. To limit cross-hybridisations, the specificity of all selected probes is checked against a large formatted database dedicated to microbial communities, *i.e.*, the EnvExBase (Environmental Expressed sequences dataBase), which is composed of 13,697,580 coding DNA sequences (CDSs) from the prokaryotic (PRO), fungal (FUN) and environmental (ENV) taxonomic divisions of the EMBL databank.
HiSpOD is provided as a web service (http://g2im.u-clermont1.fr/hispod/).

1.2.2 Probe design parameters using HiSpOD

Contrary to KASpOD, the HiSpOD program only needs a single input FASTA file containing at least one sequence. The program is able to work using both degenerate and non-degenerate

sequences. Thus, the user can submit i) consensus sequences obtained after multiple alignments of nucleic sequences, or ii) separate non-degenerate nucleic sequences.
HiSpOD offers the classical parameters for the design of effective probes, including probe length, melting temperature, complexity, and adds supplemental properties that were not considered by previous programs. Indeed the HiSpOD program performs specificity tests using BLAST ***(20)*** with an expectation value defined by the user. The identity percentage threshold and the maximal continuous stretch of nucleotide between the probe and a non-target sequence to detect putative cross-hybridisations are also user-defined parameters. In order to facilitate probe selection, cross-hybridisation results are then clustered using a single-linkage method implemented in BLASTCLUST ***(20)***. The user can define the clustering identity percentage and length thresholds.

1.2.3 Probe design results using HiSpOD

For each sequence given in the previously submitted FASTA file, two results file will be generated through HiSpOD: a .probes file providing the designed probes in FASTA format, and a .result file that summarizes the design results. The second file contains the probe sequence, its position on the sequence and the clustered cross-hybridisation results.

2. Materials

Both KASpOD and HiSpOD algorithms are provided through a web-service, an internet connected computer is therefore needed.

2.1 KASpOD requirements

1. A FASTA file containing the targeted nucleic sequences (*e.g.* the 16S rDNA sequences of the *Borrelia* genus).
2. A FASTA file containing the non-targeted nucleic sequences (*e.g.* the 16S rDNA sequences of all the *Spirochaetes* phylum except *Borrelia*).

2.2 HiSpOD requirements

1. A FASTA file containing a consensus nucleic sequence (in accordance with the IUPAC nomenclature) for a gene-specific design or a FASTA file containing non-degenerate nucleic sequences for a sequence-specific design.

According to the IUPAC code, allowed characters are: A, C, G, T, R, Y, M, K, W, S, B, D, H, V and N. HiSpOD specificity tests are performed against a comprehensive CDS database, users are therefore recommended to use CDSs as input for the probe design.

Consensus sequences can be obtained using the following strategy:

- Multiple sequence alignment using tools such as ClustalW2 ***(21)*** or Muscle ***(22)***
- Consensus creation using stand-alone software such as Seaview ***(23)***, web-services (*e.g.* Consensus Maker http://www.hiv.lanl.gov/content/sequence/CONSENSUS/consensus.html) or custom scripts.

3. Methods

3.1 How to define group-specific probes using KASpOD

This section summarizes the steps the user has to go through to perform a group-specific probe design using KASpOD.

1. Connect to the KASpOD web service.

    1.1 Go to the KASpOD website: http://g2im.u-clermont1.fr/kaspod/

    1.2 Create an account (“Create account” tab on the left menu) or log in using your login and password information.

2. Start a new probe design job.

    2.1 Click on the “New Job” tab on the left menu.

3. Select input files. See Note 1 for further details about the input data.

    3.1. Browse the target sequences file in FASTA format using the “Browse” button.

    3.2. Browse the non-target sequences file in FASTA format using the “Browse” button.

4. Customize design parameters (**Fig. 3**).

    4.1. Oligonucleotide length: Set the probe length from 18 to 31 nucleotides. [Default: 25].

4.2. Edit distance: Set the maximal number of differences (mismatches/gaps) allowed between the probe and its target or non-target. [Default: 2].

5. Launch the design.

5.1. Press the “Launch” button.

6. Wait for the server to finish processing the query. The status of the processing can be seen clicking on the “Running jobs” tab on the left menu. Job status could be: “running” if the job is currently running on a CPU node, “queue” that means that the job will be launched as soon as a CPU node is free, or “waiting” if the job has not been submitted yet to the queue management system. Computation times range from hours to days depending on the size of both the target and the non-target sequences files.

7. Download the results. Once the processing has completed the job will appear in the “Old jobs” tab with a “done” status (**Fig. 4**). Otherwise, if the probe design encountered an error, the status will be “failed”.

7.1. Click on the “Old jobs” tab on the left menu.

7.2. Click on the green arrow to download the .csv result file.

8. Load the data file. Double-click on the downloaded file to open it in MS Excel or equivalent. The table can be sorted by decreasing coverage to help the selection of the best probes. We strongly recommend the users to select multiple probes per group in order to avoid misleading interpretation of hybridisation data (e.g. cross-hybridisations, secondary structures, etc.)

3.2 How to define protein-coding gene-specific probes using HiSpOD

This section summarizes the different steps to perform a protein-coding gene-specific probe design using HiSpOD.

1. Connect to the HiSpOD web service.

1.1 Go to the HiSpOD website: http://g2im.u-clermont1.fr/hispod/

1.2 Create an account (“Create Account” tab on the left menu) or log in using your login and password information.

2. Start a new probe design job.

2.1 Click on the “New design” tab on the left menu.

3. Customize design parameters (**Fig. 5**).

3.1. “Oligonucleotide generation parameters”

3.1.1. Oligonucleotide length: Set the probe length from 18 to 120 nucleotides in the first box. [Default: 50 nucleotides].

3.1.2. Melting temperature range: Set the T_m range (in Celsius degrees) in the next two boxes. Melting temperature is computed using the following formula in which $[Na^+]$ is fixed by default at 0.5M:

$T_m = 79.8 + 18.5 \times \log_{10}([Na^+]) + 58.4 \times (yG+zC)/(wA+xT+yG+zC) + 11.8 \times (yG+zC)^2/(wA+xT+yG+zC) - 920/(wA+xT+yG+zC)$

[Default: 64-79°C].

3.1.3. Complexity: In the following box, you can set the maximal number of successive identical nucleotides allowed in the candidate probes (from 0 to probe length). For instance, a low-complexity oligonucleotide would be "AGATGC**AAAAAAAAAA**GCTGACGTA". [Default: 10].

3.1.4. Degeneracy: Click on the checkbox to set the maximal degeneracy allowed for candidate probes (from 1 to 32). For example, the following oligonucleotide "GATGAT**Y**CGTA**H**GTAGCTA**N**CTGAC" contains three degenerate oligonucleotides according to the IUPAC code (*i.e.* Y={C,T}, H={A,T,C} and N={A,C,G,T}). The degeneracy of this oligonucleotide is therefore equal to: Degeneracy = 2×3×4 = 24. [Default: 1].

3.2. "Similarity search and cross-hybridisation parameters"

3.2.1. Expectation value: Set the expectation value threshold for the specificity test using BLAST (from 0 to 40,000). [Default: 1000].

3.2.2. Identity threshold: Set the minimal percentage identity threshold between the probe and a target to consider a cross-hybridisation (from 0 to 100). [Default: 75%].

3.2.3. Identical nucleotide stretch: Set the minimal number of successive identical nucleotides between the probe and a target to consider a cross-hybridisation (from 0 to probe length). [Default: 15 nucleotides]. Kane and colleagues ***(24)*** showed that for a given oligonucleotide any non-target sequence harbouring at least 75% identity over 50 nucleotides, or sharing at least 15 contiguous identical bases will contribute to the overall signal intensity and may therefore lead to misleading interpretation of hybridisation data.

3.3. "Cross-hybridisation results"

3.3.1. Cluster similarity threshold: Set the percentage identity threshold to cluster cross-hybridizing sequences (from 0 to 100). [Default: 90%].

3.3.2. Cluster length threshold: Click on the checkbox to set the sequence length threshold to cluster cross-hybridizing sequences (from 0 to 100). A value of 0 corresponds to a clustering without taking into account the sequence length. A value of 90 will allow clustering sequences that met the percentage identity threshold over an area covering 90% of the length of each sequence. [Default: 0].

4. Select input file. See Note 1 for further details about the input data.

4.1. Browse the target sequences file in FASTA format using the "Browse" button.

5. Launch the design.

5.1. Press the "Submit" button.

5.2. Verify your design parameters and then click on the "Launch Hispod" button.

6. Wait for the server to finish processing the query. The status of the processing job can be seen clicking on the "Old results" tab on the left menu. The running time for a non-degenerate single sequence (*e.g.* about 1,000 nucleotides) is about few hours. An increasing number of sequences and degeneracy may lead to computations over few days.

7. Download the results. Once the processing job has completed it will appear in the "Old results" tab with a "done" status.

7.1. Click on the "Old results" tab on the left menu.

7.2. Click on the "Download" link next to your job.

7.3. Click on the "Download all" link at the bottom of the page to retrieve all the files in a compressed archive (.tar.gz extension). Otherwise the user can right-click and select "Save the target as…" on each file you want to download.

8. Load the data file. Double-click on the downloaded .result file to open it in a text editor. Results are sorted by decreasing number of cross-hybridisations. We strongly recommend the users to select multiple probes per group in order to avoid misleading interpretation of hybridisation data (*e.g.* cross-hybridisations, secondary structures, *etc*.). The position field of the .result file allow the users selecting probes from different regions of the gene

4. Notes

1. Problems related to input data: The most common source of problems with running KASpOD or HiSpOD is problems with input data:

1.1. Please make sure that the data is in FASTA formatted plain text files. Notice that the file must be a text-only file, either a .txt or .fasta file for instance (an otherwise correctly

formatted FASTA file within a MS-Word document will NOT work). NO .doc, .docx, .pdf, .rtf or .odt extensions allowed.

1.2. For the KASpOD program, please make sure that the first file contains only the targeted sequences. The second file should contain only the non-targeted sequences. Presence of a same sequence in both files will lead to misleading probe design results.

1.3. For the HiSpOD program, please make sure that the input file contains the sequences of the genes which should be targeted. Submitting a file with a single large DNA sequence representing an entire microbial genome will not work. HiSpOD defined gene-specific probes, for comments on how to design a chromosomal tiling array please see ***(25)***.

1.4. Please make sure that the input sequences are of a sufficient length. Entries that are shorter than the minimum probe length will be discarded.

2. Large datasets problems. Both web-services are not dedicated to very large datasets.

2.1. Input files size is limited to 16MB for the KASpOD program and 2MB for HiSpOD.

2.2. Computations are distributed on a cluster (135 CPUs) and each job cannot process more than 30 days. For large scale computations feel free to contact us.

2.2.1. Since the KASpOD web-service is not suitable for large datasets, a stand-alone version of KASpOD is available for download (64-bits GNU/Linux version). Registered users can download it through the "About" tab on the left menu.

References

1. S.M. Tiquia, L. Wu, S.C. Chong, et al. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples, BioTechniques. 36, 664–70– 672– 674–5.
2. L.A. Marcelino, V. Backman, A. Donaldson, et al. (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data, Proceedings of the National Academy of Sciences of the United States of America. 103, 13629–13634.
3. E. Dugat-Bony, M. Missaoui, E. Peyretaillade, et al. (2011) HiSpOD: probe design for functional DNA microarrays, Bioinformatics (Oxford, England). 27, 641–648.
4. A. Loy and L. Bodrossy (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays, Clinica chimica acta; international journal of clinical chemistry. 363, 106–119.
5. M. Wagner, H. Smidt, A. Loy, et al. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions, Microbial Ecology. 53, 498–506.
6. E. Dugat-Bony, E. Peyretaillade, N. Parisot, et al. (2012) Detecting unknown sequences with DNA microarrays: explorative probe design strategies, Environmental Microbiology. 14, 356–371.
7. J. Zhou (2003) Microarrays for bacterial detection and microbial community analysis, Current Opinion in Microbiology. 6, 288–294.
8. A. Loy, A. Lehner, N. Lee, et al. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment, Applied and Environmental Microbiology. 68, 5064–5081.
9. K.H. Wilson, W.J. Wilson, J.L. Radosevich, et al. (2002) High-density microarray of small-subunit ribosomal DNA probes, Applied and Environmental Microbiology. 68, 2535–2541.

10. E.L. Brodie, T.Z. DeSantis, D.C. Joyner, et al. (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation, Applied and Environmental Microbiology. 72, 6288–6298.
11. C. Palmer, E.M. Bik, M.B. Eisen, et al. (2006) Rapid quantitative profiling of complex microbial populations, Nucleic Acids Research. 34, e5.
12. E.L. Brodie, T.Z. DeSantis, J.P.M. Parker, et al. (2007) Urban aerosols harbor diverse and dynamic bacterial populations, Proceedings of the National Academy of Sciences of the United States of America. 104, 299–304.
13. T.Z. DeSantis, E.L. Brodie, J.P. Moberg, et al. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment, Microbial Ecology. 53, 371–383.
14. T.C. Hazen, E.A. Dubinsky, T.Z. DeSantis, et al. (2010) Deep-sea oil plume enriches indigenous oil-degrading bacteria, Science. 330, 204–208.
15. T.C. Hazen, A.M. Rocha, and S.M. Techtmann (2013) Advances in monitoring environmental microbes, Current opinion in biotechnology. 24, 526–533.
16. C. Militon, S. Rimour, M. Missaoui, et al. (2007) PhylArray: phylogenetic probe design algorithm for microarray, Bioinformatics (Oxford, England). 23, 2550–2557.
17. N. Parisot, J. Denonfoux, E. Dugat-Bony, et al. (2012) KASpOD--a web service for highly specific and explorative oligonucleotide design, Bioinformatics (Oxford, England). 28, 3161–3162.
18. Z. He, J.D. Van Nostrand, L. Wu, et al. (2008) Development and application of functional gene arrays for microbial community analysis, Transactions of Nonferrous Metals Society of China. 18, 1319–1327.
19. S. Lemoine, F. Combes, and S. Le Crom (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge, Nucleic Acids Research. 37, 1726–1739.
20. S.F. Altschul, W. Gish, W. Miller, et al. (1990) Basic local alignment search tool, Journal of Molecular Biology. 215, 403–410.
21. J.D. Thompson, D.G. Higgins, and T.J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research. 22, 4673–4680.
22. R.C. Edgar (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research.
23. M. Gouy, S. Guindon, and O. Gascuel (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building, Molecular Biology and Evolution.
24. M.D. Kane, T.A. Jatkoe, C.R. Stumpf, et al. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays, Nucleic Acids Research. 28, 4552–4557.
25. N. Parisot, J. Denonfoux, E. Dugat-Bony, et al. Software Tools for the Selection of Oligonucleotide Probes for Microarrays, Microarrays: Current Technology, Innovations and Applications, Horizon Scientific Press.
26. W. Ludwig, O. Strunk, R. Westram, et al. (2004) ARB: a software environment for sequence data, Nucleic Acids Research. 32, 1363–1371.
27. K.C. Bader, C. Grothoff, and H. Meier (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets, Bioinformatics (Oxford, England). 27, 1546–1554.
28. Q. Tu, Z. He, Y. Deng, et al. (2013) Strain/Species-Specific Probe Design for Microbial Identification Microarrays, Applied and Environmental Microbiology.
29. M. Severgnini, P. Cremonesi, C. Consolandi, et al. (2009) ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design, Nucleic Acids Research. 37, e109.
30. H.L. Fredrickson, E.J. Perkins, T.S. Bridges, et al. (2001) Towards environmental toxicogenomics -- development of a flow-through, high-density DNA hybridization array and its application to ecotoxicity assessment, The Science of the total environment. 274, 137–149.
31. S. Feng and E.R. Tillier (2007) A fast and flexible approach to oligonucleotide probe design for genomes and gene families, Bioinformatics (Oxford, England). 23, 1195–1202.
32. Z. Bozdech, J. Zhu, M.P. Joachimiak, et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray, Genome Biology. 4, R9.
33. L. Ilie, H. Mohamadi, G.B. Golding, et al. (2013) BOND: Basic OligoNucleotide Design, BMC Bioinformatics. 14, 69–69.
34. X. Li, Z. He, and J. Zhou (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation, Nucleic Acids Research. 33, 6114–6123.
35. W. Li and X. Ying (2006) Mprobe 2.0: computer-aided probe design for oligonucleotide microarray, Applied bioinformatics. 5, 181–186.
36. J.-M. Rouillard, M. Zuker, and E. Gulari (2003) OligoArray 2.0: design of oligonucleotide probes for

DNA microarrays using a thermodynamic approach, Nucleic Acids Research. 31, 3057–3062.
37. X. Wang and B. Seed (2003) Selection of oligonucleotide probes for protein coding sequences, Bioinformatics (Oxford, England). 19, 796–802.
38. R. Wernersson and H.B. Nielsen (2005) OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes, Nucleic Acids Research. 33, W611–5.
39. D. Xu, G. Li, L. Wu, et al. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis, Bioinformatics (Oxford, England). 18, 1432–1437.
40. S.-H. Chen, C.-Z. Lo, S.-Y. Su, et al. (2010) UPS 2.0: unique probe selector for probe design and oligonucleotide microarrays at the pangenomic/genomic level, BMC Genomics. 11 Suppl 4, S6.
41. E.C. Rouchka, A. Khalyfa, and N.G.F. Cooper (2005) MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays, BMC Bioinformatics. 6, 175.
42. E. Talla, F. Tekaia, L. Brino, et al. (2003) A novel design of whole-genome microarray probes for Saccharomyces cerevisiae which minimizes cross-hybridization, BMC Genomics. 4, 38.
43. H.-H. Chou, A.-P. Hsia, D.L. Mooney, et al. (2004) Picky: oligo microarray design for large genomes, Bioinformatics (Oxford, England). 20, 2893–2902.
44. F. Li and G.D. Stormo (2001) Selection of optimal DNA oligos for gene expression arrays, Bioinformatics (Oxford, England). 17, 1067–1076.
45. Y.S. Dufour, G.E. Wesenberg, A.J. Tritt, et al. (2010) chipD: a web tool to design oligonucleotide probes for high-density tiling arrays, Nucleic Acids Research. 38, W321–5.
46. E. Ryder, R. Jackson, A. Ferguson-Smith, et al. (2006) MAMMOT--a set of tools for the design, management and visualization of genomic tiling arrays, Bioinformatics (Oxford, England). 22, 883–884.
47. V.C. Patel, K. Mondal, A.C. Shetty, et al. (2010) Microarray oligonucleotide probe designer (MOPeD): A web service, Open access bioinformatics. 2, 145–155.
48. P. Bertone, V. Trifonov, J.S. Rozowsky, et al. (2006) Design optimization methods for genomic DNA tiling arrays, Genome Research. 16, 271–281.
49. A.M. Phillippy, X. Deng, W. Zhang, et al. (2009) Efficient oligonucleotide probe selection for pan-genomic tiling arrays, BMC Bioinformatics. 10, 293.
50. L. Jourdren, A. Duclos, C. Brion, et al. (2010) Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments, Nucleic Acids Research. 38, e117.

Acknowledgements

This work was supported by the French “Direction Générale de l’Armement” (DGA) and the programme Investissements d’avenir AMI 2011 VALTEX.

Figure captions

Figure 1. Overview of the KASpOD algorithm.

Figure 2. Overview of the HiSpOD program workflow.

Figure 3. Screenshot of the KASpOD “new design” tab.

Figure 4. Screenshot of the KASpOD “old jobs” tab.

Figure 5. Screenshot of the HiSpOD “new design” tab.

Table captions

Table 1. Comparison of oligonucleotide probe design software.

Tables

Table 1. Comparison of oligonucleotide probe design software.

Software	Reference	Application	Availability	URL
ARB (v 5.5)	*(26)*	POA	Downloadable, standalone GUI (L, M)	http://www.arb-home.de/
CaSSiS (v 0.5.0)	*(27)*	POA	Downloadable, command-line (L)	http://cassis.in.tum.de
PhylArray	*(16)*	POA	Web interface	http://g2im.u-clermont1.fr/serimour/phylarray
SSPD	*(28)*	POA	Web interface	http://ieg.ou.edu/SSPD/
ORMA	*(29)*	POA, FGA	Matlab Script	Upon request
KASpOD	*(17)*	POA, FGA, WGA-ORF	Web interface or command-line (L)	http://g2im.u-clermont1.fr/kaspod/
DEODAS (v 0.1.0)	*(30)*	FGA	Downloadable, GUI (L)	http://deodas.sourceforge.net/
ProDesign	*(31)*	FGA	Web interface	http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html
ArrayOligoSelector (v 3.8.4)	*(32)*	FGA, WGA-ORF	Downloadable, command-line (L)	http://arrayoligosel.sourceforge.net
BOND	*(33)*	FGA, WGA-ORF	Downloadable, command-line (L, W, M)	www.csd.uwo.ca/~ilie/BOND/
CommOligo (v 2.0)	*(34)*	FGA, WGA-ORF	Downloadable, standalone GUI (W)	http://ieg.ou.edu/software.htm
HiSpOD	*(3)*	FGA, WGA-ORF	Web interface	http://g2im.u-clermont1.fr/hispod/
MProbe (v 2.0)	*(35)*	FGA, WGA-ORF	Downloadable, GUI (W)	http://www.biosun.org.cn/mprobe/
OligoArray (v 2.1)	*(36)*	FGA, WGA-ORF	Downloadable, command-line (L)	http://berry.engin.umich.edu/oligoarray2_1/
OligoPicker (v 2.3.2)	*(37)*1	FGA, WGA-ORF	Downloadable, command-line (L)	http://pga.mgh.harvard.edu/oligopicker/
OligoWiz (v 2.3.0)	*(38)*	FGA, WGA-ORF	Downloadable client program, GUI (L, W, M)	http://www.cbs.dtu.dk/services/OligoWiz2
PRIMEGENS (v 2.0)	*(39)*	FGA, WGA-ORF	Web interface or command-line standalone (L, W)	http://primegens.org/
UPS 2.0	*(40)*	FGA, WGA-ORF	Web interface	http://array.iis.sinica.edu.tw/ups/
Mprime	*(41)*	WGA-ORF	Web interface	http://kbrin.a-bldg.louisville.edu/Tools/OligoDesign/MPrime.html
OliD	*(42)*	WGA-ORF	Downloadable, command line (L)	Upon request
PICKY (v 2.2)	*(43)*	WGA-ORF	Downloadable, standalone GUI (L, W, M)	http://www.complex.iastate.edu./download/Picky/index.html
ProbeSelect	*(44)*	WGA-ORF	Available upon request, command line (L)	http://stormo.wustl.edu/src/probeselect-src.tar
ChipD	*(45)*	WGA-tiling	Web interface	http://chipd.uwbacter.org/
MAMMOT (v 1.21)	*(46)*	WGA-tiling	Downloadable, local server	http://www.mammot.org.uk/
MOPeD	*(47)*	WGA-tiling	Web interface	http://moped.genetics.emory.edu/newdesign.html
OligoTiler	*(48)*	WGA-tiling	Web interface	http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi
PanArray (v 1.0)	*(49)*	WGA-tiling	Available upon request, command line (L)	Upon request
Teolenn (v 2.0.1)	*(50)*	WGA-tiling	Downloadable, command line (L)	http://transcriptome.ens.fr/teolenn/

POA: phylogenetic oligonucleotide array. FGA: functional gene array. WGA-ORF: open reading-frame oriented whole-genome array. WGA-tiling: tiling whole-genome array. GUI: graphical user interface. L: Linux. M: MacOS. W: Windows.

Figure 1

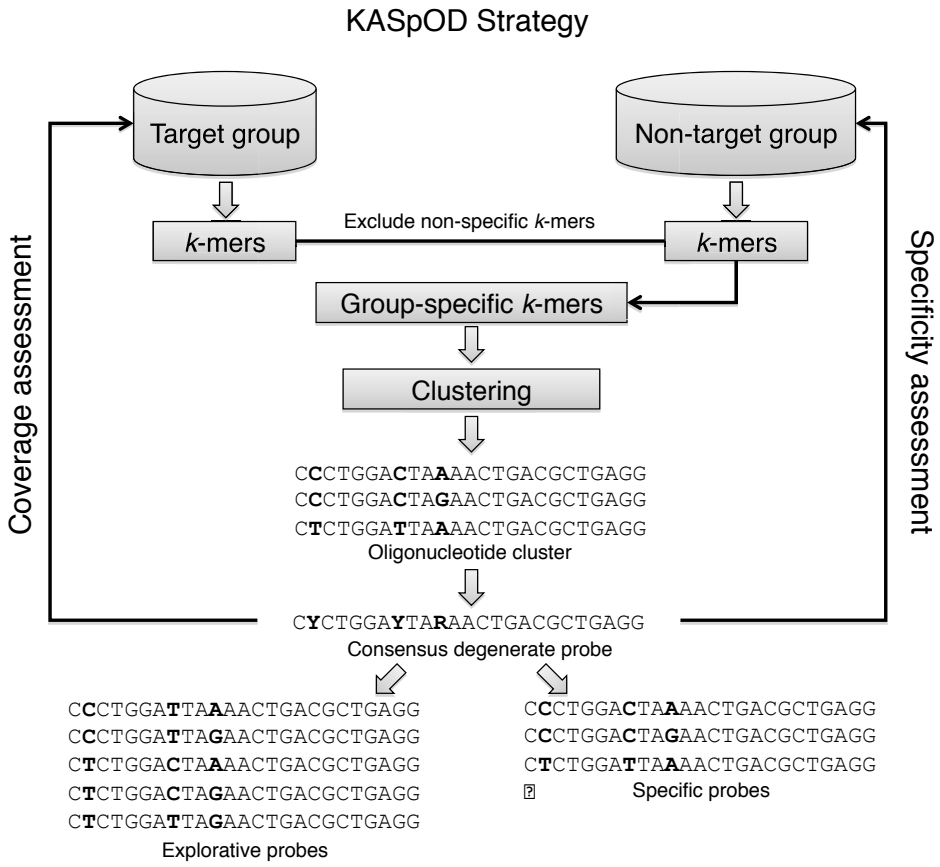

KASpOD Strategy
Target group
Non-target group
k-mers
Exclude non-specific k-mers
k-mers
Group-specific k-mers
Clustering
CCCTGGACTAAAACTGACGCTGAGG
CCCTGGACTAGAACTGACGCTGAGG
CTCTGGATTAAAACTGACGCTGAGG
Oligonucleotide cluster
CYCTGGAYTARAACTGACGCTGAGG
Consensus degenerate probe
CCCTGGATTAAAACTGACGCTGAGG
CCCTGGATTAGAACTGACGCTGAGG
CTCTGGACTAAAACTGACGCTGAGG
CTCTGGACTAGAACTGACGCTGAGG
CTCTGGATTAGAACTGACGCTGAGG
Explorative probes
CCCTGGACTAAAACTGACGCTGAGG
CCCTGGACTAGAACTGACGCTGAGG
CTCTGGATTAAAACTGACGCTGAGG
Specific probes
Coverage assessment
Specificity assessment

Figure 2

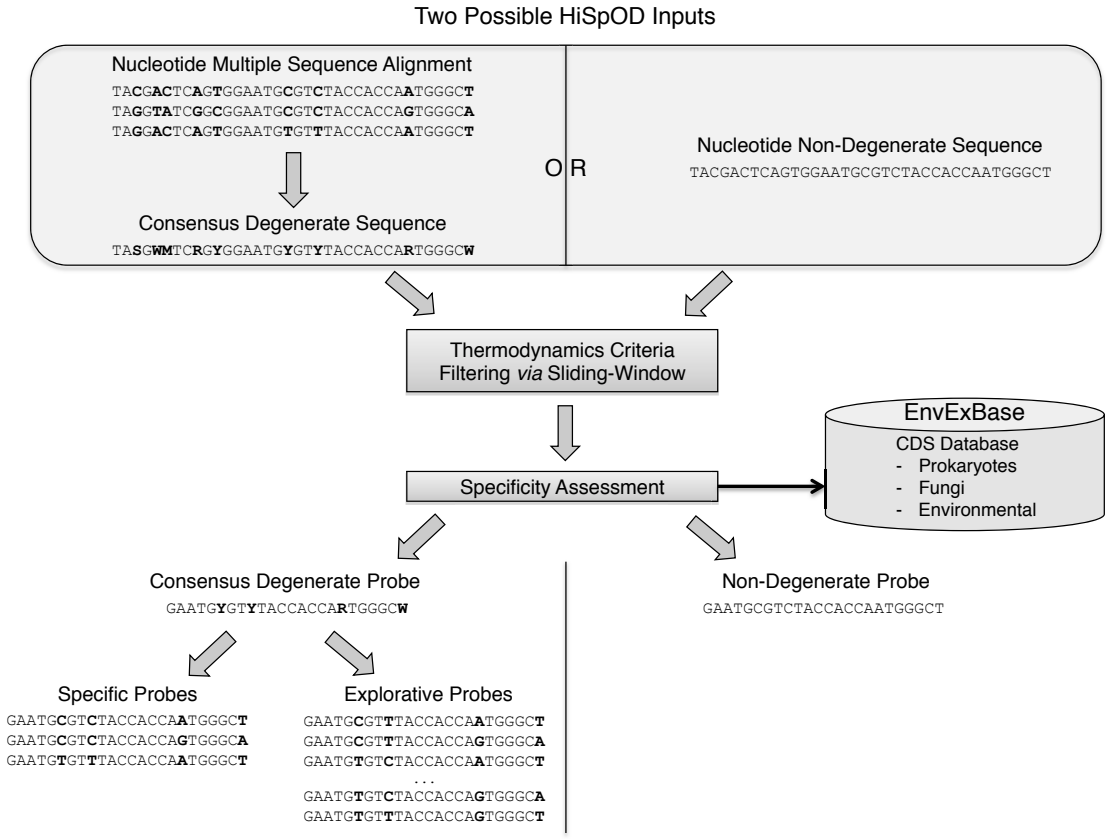

Two Possible HiSpOD Inputs
Nucleotide Multiple Sequence Alignment
TACGACTCAGTGGAATGCGTCTACCACCAATGGGCT
TAGGTATCGGCGGAATGCGTCTACCACCAGTGGGCA
TAGGACTCAGTGGAATGTGTTTACCACCAATGGGCT
Consensus Degenerate Sequence
TASGWMTCRGYGGAATGYGTYTACCACCARTGGGCW
O R
Nucleotide Non-Degenerate Sequence
TACGACTCAGTGGAATGCGTCTACCACCAATGGGCT
Thermodynamics Criteria Filtering via Sliding-Window
EnvExBase
CDS Database
- Prokaryotes
- Fungi
- Environmental
Specificity Assessment
Consensus Degenerate Probe
GAATGYGTYTACCACCARTGGGCW
Non-Degenerate Probe
GAATGCGTCTACCACCAATGGGCT
Specific Probes
GAATGCGTCTACCACCAATGGGCT
GAATGCGTCTACCACCAGTGGGCA
GAATGTGTTTACCACCAATGGGCT
Explorative Probes
GAATGCGTTTACCACCAATGGGCT
GAATGCGTTTACCACCAGTGGGCA
GAATGTGTCTACCACCAATGGGCT
...
GAATGTGTCTACCACCAGTGGGCA
GAATGTGTTTACCACCAGTGGGCT

Figure 3

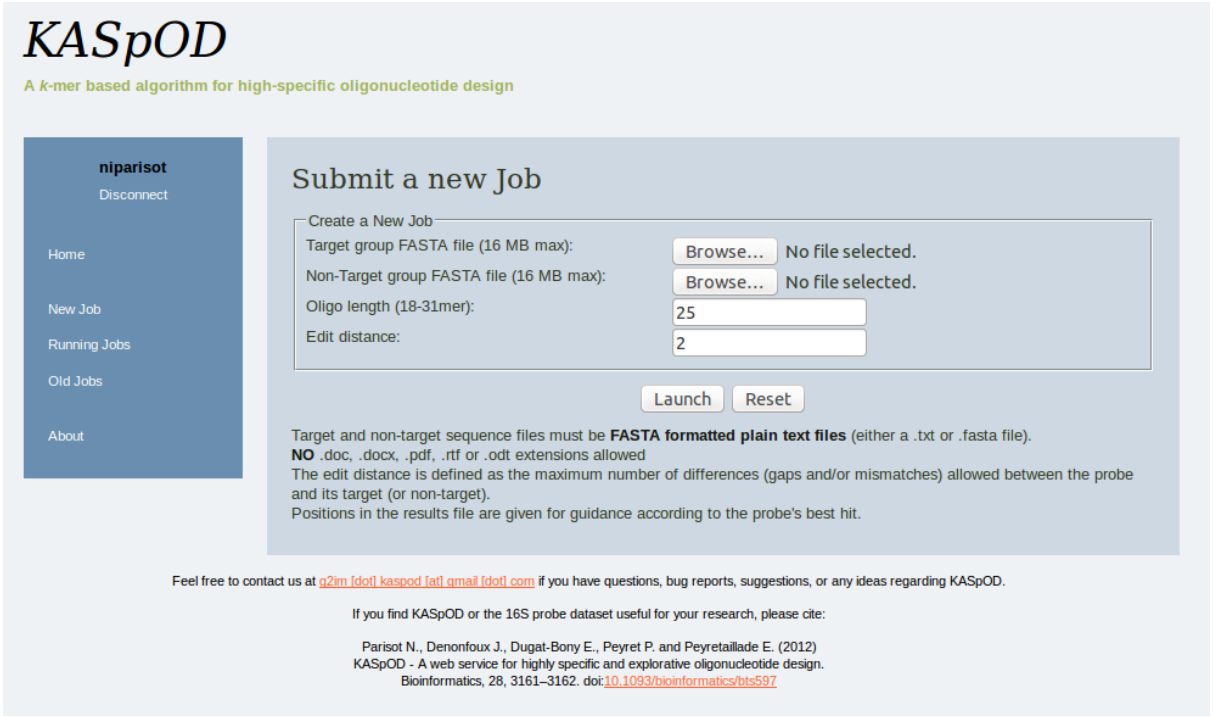


Figure 4

Figure 5


HiSpOD
High Specific Oligo Design:
probe design for functional microarrays
Menu
03 Dec 13 -11:12:44
Disconnect
Description
New design
Old results
Linker
Probe design parameters
(1) Oligonucleotide generation parameters
Probe length : 50
Tm min : 64
Tm max : 79
Complexity : 10
Degeneracy :
(2)&(3) Similarity Search & cross-hybridization parameters
Expected value : 1000
Identity % : 75
Similarity stretch : 15
(4) Cross-Hybridization results
Cluster similarity : 90
Cluster length :
(5) Data parameters
Reference database : EnvExBase
Input file (FASTA) : Browse... No file selected.
Submit
Reset

2.4 Discussion

A l'heure actuelle, de nombreux logiciels de détermination de sondes pour biopuces ADN sont disponibles, mais peu d'entre eux sont appliqués pour des études environnementales (Lemoine *et al.* 2009 ; Dugat-Bony *et al.* 2012b ; Parisot *et al.* 2014). Le développement de nouveaux logiciels, proposant des nouvelles stratégies permettant de répondre aux exigences et aux contraintes de l'écologie microbienne, apparaît donc nécessaire. C'est avec cet objectif que le logiciel KASpOD a été développé. Il offre de nouvelles opportunités en combinant les atouts de logiciels dédiés pour la détermination de sondes pour POA comme PhylArray (Militon *et al.* 2007) et PhylGrid 2.0 (Jaziri *et al.* 2014b) ou pour FGA comme HiSpOD (Dugat-Bony *et al.* 2011) et Metabolic Design (Terrat *et al.* 2010), qui intègrent le caractère exploratoire des sondes tout en optimisant la recherche des hybridations croisées potentielles. Cependant, la stratégie développée au travers de KASpOD est différente de celles proposées pour les autres logiciels. En effet, la détermination des sondes ne se fait pas à partir du résultat d'alignements multiples de séquences, limitants pour des jeux de données importants, mais à partir de la recherche de motifs nucléiques (ou *k*-mers) spécifiques des séquences ciblées. Ces *k*-mers sont recherchés et extraits des séquences cibles données en entrée en utilisant l'outil Jellyfish (Marcais & Kingsford 2011). Une telle approche permet de fortement réduire les temps de calcul et l'usage de mémoire. Par ailleurs, contrairement aux autres logiciels, les différents tests de couverture et de spécificité n'utilisent pas l'approche BLAST, mais PatMaN (*Pattern Matching in Nucleotide databases*) (Prüfer *et al.* 2008) qui est capable de rechercher rapidement et de manière exhaustive toutes les occurrences, exactes ou non, de courtes séquences nucléiques au sein d'un large jeu de données de séquences. En effet, PatMan peut récupérer les occurrences non exactes en permettant à l'utilisateur de fixer un nombre maximal de différences (*gaps* ou mésappariements) autorisées entre la séquence testée et la séquence de la base de données. Le logiciel BLAST, quant à lui, n'a pas été développé pour cette application et se base sur la recherche de mots exacts d'au minimum 7 nucléotides. Sa sensibilité reste donc limitée pour la recherche de similarités à partir de courtes séquences comme les sondes oligonucléotidiques. De plus, cet outil est capable d'utiliser une séquence dégénérée comme requête contrairement au BLAST, qui impose une analyse par combinaison non dégénérée. L'outil PatMaN apparaît donc beaucoup plus adapté que le logiciel BLAST pour cette application.

Cependant, une limite de KASpOD est qu'il ne permet pas de définir des sondes d'une taille supérieure à 31-mers. Aussi, afin de disposer de sondes longues permettant d'augmenter la sensibilité, une alternative serait d'appliquer la stratégie GoArrays (Rimour *et al.* 2005) sur les sondes définies avec KASpOD. De même, les critères thermodynamiques comme la température de fusion (T_m) ou la formation de structures secondaires ne sont pas pris en compte pour la sélection des sondes. Une amélioration intéressante consisterait donc à intégrer ces critères à la stratégie KASpOD.

Le logiciel KASpOD se présente donc comme un nouvel outil performant pour disposer de sondes présentant une très bonne couverture, une grande spécificité et possédant le caractère exploratoire. Dans le cadre de l'écologie microbienne, et plus particulièrement de l'étude des environnements complexes, cette nouvelle stratégie de détermination de sondes présente toutes les qualités pour définir des sondes de qualité pour différentes applications allant de la PCR à la capture de gènes en passant par les biopuces ADN (POA, FGA ou WGA).

3. Développement d'une base de données de sondes oligonucléotidiques ciblant le gène ADNr 16S : PhylOPDb

3.1 Contexte

Les méthodes moléculaires reposant sur l'utilisation de sondes oligonucléotidiques comme la PCR, les biopuces ADN, le FISH ou encore la capture de gènes connaissent un essor important et permettent aujourd'hui d'aborder de manière ciblée les problématiques d'écologie microbienne. Avec elles, de nombreux logiciels de détermination de sondes ont également vu le jour (Parisot *et al.* 2014). Néanmoins, bien que de nombreuses sondes oligonucléotidiques aient été utilisées, il reste aujourd'hui très difficile d'accéder à des collections de sondes, si ce n'est au travers d'études bibliographiques fastidieuses.

L'apport des techniques de séquençage haut-débit a permis l'étude intensive des communautés microbiennes *via* l'analyse de biomarqueurs phylogénétiques comme le gène ADNr 16S. Une croissance exponentielle du nombre de séquences d'ADNr 16S dans les bases de données internationales s'est alors opérée. Cependant, il n'existe aujourd'hui que peu de bases de données répertoriant des sondes oligonucléotidiques ciblant le gène ADNr 16S. Il est possible de citer la base de données « *Oligonucleotide Probe Database* » (OPD) (Alm *et al.* 1996) développée en 1996 et regroupant 96 amorces ou sondes oligonucléotidiques dirigées vers les séquences codant la petite ou la grande sous-unité de l'ARN ribosomique. Cette base de données n'est hélas plus maintenue depuis 1997. Plus récemment, la base de données « probeBase » (Loy *et al.* 2007) a été mise au point et compte aujourd'hui 2 788 sondes (24/07/14).

3.2 Objectif

L'objectif de ce travail a donc consisté à mettre en place une base de données de sondes oligonucléotidiques ciblant le gène ADNr 16S. Une telle structure de données permettrait de rendre accessibles à la communauté scientifique, des signatures oligonucléotidiques utilisables pour différentes applications : biopuces ADN, PCR, FISH, capture de gènes etc.

Ainsi, grâce aux deux algorithmes précédemment décrits, PhylGrid 2.0 (Jaziri *et al.* 2014b) et KASpOD (Parisot *et al.* 2012), deux jeux de sondes oligonucléotidiques ont pu être obtenus. PhylGrid 2.0 a utilisé une base de données propriétaire de séquences d'ADNr 16S

composée de 66 075 séquences définissant 2069 genres procaryotes. A partir de ces séquences, 19 874 sondes de 25 nucléotides, ciblant l'ensemble des genres, ont été sélectionnées. L'algorithme KASpOD a, quant à lui, été appliqué sur un sous-ensemble de 252 183 séquences de haute-qualité représentant 1295 genres procaryotes issus de la base de données Greengenes (McDonald *et al.* 2012). 54 129 sondes supplémentaires de 25-mers ont alors pu être sélectionnées. Au final, ce sont 74 003 sondes, alliant sensibilité, spécificité et caractère exploratoire, qui ont permis de constituer la base de données de sondes oligonucléotidiques, ciblant le gène ADNr 16S, la plus exhaustive à l'heure actuelle. Cette base de données a été nommée PhylOPDb pour « *Phylogenetic Oligonucleotide Probe Database* ».

3.3 Principaux résultats

Le travail réalisé a donc permis de proposer une nouvelle base de données de sondes oligonucléotidiques ciblant le gène ADNr 16S. La base de données PhylOPDb, consultable *via* une interface web (http://g2im.u-clermont1.fr/phylopdb/), a donné lieu à une publication dans le journal « *Database* ».

L'interface web, développée en utilisant les langages PHP, HTML5, CSS3, JavaScript, jQuery, JSON et MySQL, permet une consultation aisée et une interrogation efficace de la base de données. Il existe différents moyens d'accéder aux sondes oligonucléotidiques d'intérêt au sein de PhylOPDb: i) par une navigation hiérarchique grâce à la taxonomie EMBL, ii) par une recherche textuelle ou enfin iii) par une recherche avancée suivant différents critères. Ainsi, la navigation hiérarchique permet à l'utilisateur de parcourir les différents niveaux taxonomiques (*i.e.* genre, famille, classe, ordre, phylum et domaine) afin de sélectionner le taxon de son choix. Lorsqu'un groupe taxonomique est sélectionné, les sondes ciblant les genres appartenant à ce groupe sont alors affichées. L'utilisateur a ensuite la possibilité de télécharger les résultats dans un format CSV ou FASTA. Pour chaque sonde oligonucléotidique, différentes informations sont disponibles : l'identifiant de la sonde, le genre ciblé, sa séquence, sa longueur, la séquence de la sonde dégénérée dont elle provient, sa dégénérescence, sa couverture, sa position, les potentielles hybridations croisées qu'elle peut produire, son T_m et l'outil de *design* utilisé pour déterminer cette sonde. La recherche textuelle permet à l'utilisateur de rapidement identifier les sondes qui l'intéresse grâce à la saisie d'une chaine de caractères qui sera recherchée dans les différents champs. Cette chaine de caractère peut alors correspondre à une partie de la séquence de la sonde, un nom de genre

ou encore un des deux logiciels utilisés pour la détermination de sondes (PhylGrid et KASpOD). La recherche avancée, quant à elle, permet à l'utilisateur d'utiliser plusieurs filtres pour récupérer les sondes d'intérêt. On peut citer le nom du genre ciblé, l'outil de *design*, la gamme de T_m (T_m minimal et T_m maximal), le nombre maximum d'hybridations croisées autorisées ou le taux de couverture minimal de la sonde. Pour exemple, l'utilisateur peut, s'il le souhaite, sélectionner rapidement les sondes déterminées avec l'outil KASpOD, ciblant au moins 50% des séquences du genre *Abiotrophia*, sans hybridations croisées et ayant un T_m compris entre 41 et 66°C.

Article n°5

PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification.

DATABASE
The Journal of Biological Databases and Curation

Database, 2014, 1–7
doi: 10.1093/database/bau036
Original article

Original article

PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification

Faouzi Jaziri[1,2,†], Nicolas Parisot[1,3,†], Anis Abid[2], Jérémie Denonfoux[1,3], Céline Ribière[1], Cyrielle Gasc[1], Delphine Boucher[1], Jean-François Brugère[1], Antoine Mahul[4], David R.C. Hill[2], Eric Peyretaillade[1] and Pierre Peyret[1,*]

[1]Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, F-63001 Clermont-Ferrand, France, [2]UMR CNRS 6158, ISIMA/LIMOS, Clermont Université, Université Blaise Pascal, F-63173 Aubière, France, [3]CNRS, UMR 6023, LMGE, F-63171 Aubière, France and [4]Clermont Université, CRRI, F-63177 Aubière, France

*Corresponding author: Tel: +33 473 178 308; Fax: +33 473 178 392; Email: pierre.peyret@udamail.fr

[†]These authors contributed equally to this work.

Citation details: Jaziri,F., Parisot,N., Abid,A. *et al.* PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification. *Database* (2014) Vol. 2014: article ID bau036; doi:10.1093/database/bau036

Received 21 October 2013; Revised 28 March 2014; Accepted 11 April 2014

Abstract

In recent years, high-throughput molecular tools have led to an exponential growth of available 16S rRNA gene sequences. Incorporating such data, molecular tools based on target-probe hybridization were developed to monitor microbial communities within complex environments. Unfortunately, only a few 16S rRNA gene-targeted probe collections were described. Here, we present PhylOPDb, an online resource for a comprehensive phylogenetic oligonucleotide probe database. PhylOPDb provides a convivial and easy-to-use web interface to browse both regular and explorative 16S rRNA-targeted probes. Such probes set or subset could be used to globally monitor known and unknown prokaryotic communities through various techniques including DNA microarrays, polymerase chain reaction (PCR), fluorescent *in situ* hybridization (FISH), targeted gene capture or *in silico* rapid sequence identification. PhylOPDb contains 74 003 25-mer probes targeting 2178 genera including *Bacteria* and *Archaea*.
Database URL: http://g2im.u-clermont1.fr/phylopdb/

Background

Prokaryotes are the most important and diverse group of organisms, widely distributed across almost all environmental habitats, even the most extreme, and involved in various ecological and environmental processes. However, because of this tremendous diversity and the technological limits such as our inability to culture the vast majority of microorganisms (1), our current vision of the microbial world is still incomplete. Thus, the comprehension of prokaryote diversity, abundance and dynamics remains a

major challenge of microbial ecology. To overcome the limitations of the culture-based methods, some molecular tools were therefore developed to survey prokaryotic communities (2) such as polymerase chain reaction (PCR)-based DNA fingerprints, fluorescent *in situ* hybridization (FISH) or clone libraries sequencing. Over the past decades, most promising high-throughput approaches were developed including DNA microarrays and next-generation sequencing that can also be coupled to gene capture (3).

Targeting the small subunit (SSU) ribosomal RNA gene, i.e. 16S rRNA gene, is particularly well adapted to survey prokaryotic communities in complex environments, as it contains highly conserved and variable moieties permitting reliable and detailed bacterial classification. Moreover, the advent of many PCR-based approaches, as well as sequencing projects, has led to the explosion of 16S rRNA gene sequences now available in major specialized sequence repositories, such as Greengenes (4), SILVA (5) and RDP (6). Taking into account this amount of data, high-throughput tools using the SSU rRNA biomarker such as phylogenetic oligonucleotide arrays (POAs) have been developed. Several tools were therefore proposed to select phylogenetic probes such as PRIMROSE (7), ARB PROBE_DESIGN (8), ORMA (9) or CaSSiS (10, 11). Unfortunately, most of these programs are not well-suited for large-scale probe design. Designing probes for a large group of sequences requires considerable computational resources and can take up to few days for only one design. Only CaSSiS was implemented for large-scale sequence data sets. Furthermore, all of these tools allow selecting probes targeting only known microbial communities with available sequences in public environmental databases. However, it is also important to define explorative probes that can detect uncharacterized phylogenetic signatures and anticipate genetic variations (12). Currently, only four software programs allow the selection of explorative phylogenetic probes: PhylArray (13), PhylGrid (14), KASpOD (15) and MetaExploArrays (16).

Although numerous oligonucleotide probes have been reported, available collections of rRNA-targeted oligonucleotide probes are rare. The oligonucleotide probe database (OPD) (17) was proposed in 1996 to collect tested phylogenetic oligonucleotide probes. The last data set of OPD listed 96 primers and probes targeting small and large subunit rRNA. However, OPD has not been updated since 1997 and is no longer available online. More recently, 'probeBase' (18), an online resource for published rRNA-targeted oligonucleotide probes and associated information, was established in 2002. It currently includes 2788 probes (status January 2014).

Here, we present PhylOPDb ('phylogenetic oligonucleotide probe database'), a comprehensive phylogenetic OPD targeting 16S rRNA gene sequences. We used two high-throughput probe design software, PhylGrid and KASpOD, to select both regular and explorative 16S rRNA gene-targeted oligonucleotide probes. PhylOPDb is composed of 74 003 probes of 25 mer targeting 2178 genera including *Bacteria* and *Archaea*.

Database construction and development

Probe design using PhylGrid

PhylGrid (14) is a large-scale probe design software linked to the EGI grid. It is an improvement of the PhylArray algorithm presented in Militon *et al.* (13) that allows defining regular and explorative oligonucleotide probes targeting SSU rRNA genes at any phylogenetic level (Figure 1).

The PhylGrid probe design was based on a custom 16S rDNA-curated sequence database originating from the EMBL. All SSU rDNA sequences downloaded from the prokaryotic (PRO) and environmental (ENV) divisions of the EMBL nucleotide sequence database were used as a reference to build this database. First, 16S rDNA gene sequences were extracted and filtered according to their quality and size. Only sequences that met the following criteria were kept: (i) sequence length is between 1200 and 1600 nucleotides, (ii) sequence is assigned to the genus level in EMBL database, (iii) the percentage of ambiguous nucleotides is <1% and (iv) the maximum number of consecutive unknown bases must not exceed five. Then, extracted sequences were grouped at the genus taxonomic rank according to the NCBI taxonomy database. For each genus, sequence orientation of all sequences was checked using BLASTN (19) and redundancy was eliminated using BLASTCLUST (19). Finally, a K-means approach was implemented to check for badly annotated sequences that may prevent selecting specific probes for this group. Eventually, 66 075 rDNA (16S) gene sequences representing 2069 prokaryotic genera were obtained.

Using this custom 16S rRNA gene sequences database, a total of 3 553 975 degenerate probes of 25 mer were selected. The probe length of 25 nucleotides offers the balance of highest sensitivity and specificity in the presence of a complex background (20, 13). Coverage and specificity tests were performed against the input database using a BLASTN (19) allowing up to two mismatches between a probe and its target. Putative cross-hybridizations were therefore defined when a non-targeted sequence harbours at most two mismatches with a probe. The number of mismatches allowed was chosen taking into the destabilization effect on the probe–target complex and the loss of signal (21). A stringent threshold of two mismatches was set to limit putative cross-hybridizations. It should be noticed

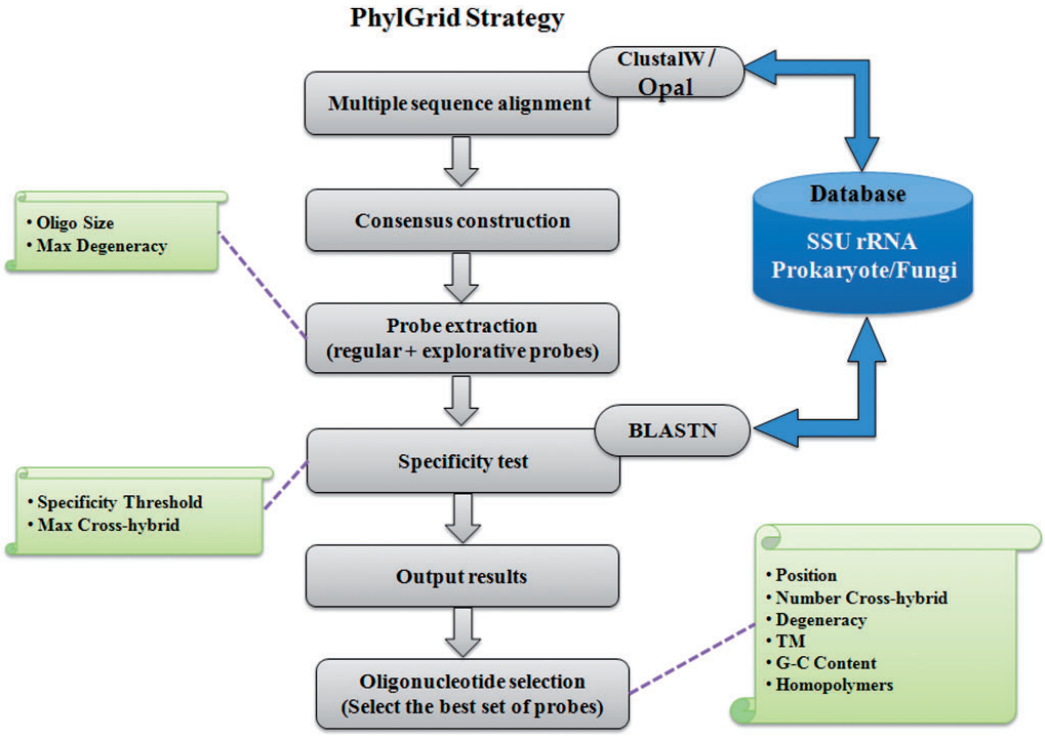


Figure 1. Overview of the PhylGrid algorithm.

that ambiguous nucleotides are considered as mismatches by BLASTN.

Five non-overlapping probes showing the best specificity and coverage were then selected for each genus. During this step, some cross-hybridizing probes may have been selected. Nevertheless, the program ensured that the simultaneous analysis of the five selected probes could not cause misleading interpretations of hybridization data. A set of 19 874 25-mer probes corresponding to 10 320 degenerate probes was obtained.

Probe design using KASpOD

KASpOD (15) is a fast *k*-mer–based software dedicated to the design of group-covering oligonucleotide probes. It allows selecting highly specific and explorative probes based on large data sets (Figure 2).

The KASpOD probe design was based on the Greengenes database (22). The May 2011 release containing 406 997 sequences was downloaded and extracted from Greengenes. Then, using a custom PERL script, only the sequences assigned to a genus were retained for further analyses. These 310 575 sequences were then sorted by genus into different FASTA files. For each genus, a clustering step was performed at a 100% identity threshold using CD-HIT (23) to remove any redundancies. Moreover, only high-quality sequences were retained: (i) sequence length >1200 nucleotides and (ii) <1% of ambiguous nucleotides. After this processing pipeline, the 16S rDNA database contained 252 250 high-quality sequences. The clustering of the whole database at high-identity thresholds (99, 98 and 97%) coupled with manual curation, allowed removing of potentially badly assigned sequences. Furthermore, some microbial genera were clustered together, as they were hardly distinguishable on the basis of their sequences. Eventually, 252 183 16S rDNA sequences were fed to KASpOD to perform the probe design.

A total of 3 242 105 degenerate candidate probes of 25 mer were designed for 1295 prokaryotic genera. The maximum number of mismatches between a probe and its target was set to two mismatches, and any ambiguous character will be counted as a match if the aligning base is one of the nucleotides represented by the ambiguity code.

The minimal probeset harbouring the best coverage and specificity was then defined. First, the non-overlapping probes were selected within the probes showing no cross-hybridizations. Subsequently, while there were some 16S rDNA sequences, which were not covered by at least three probes, the program selected additional probes with increasing numbers of cross-hybridizations. During this step, the program ensured that no more than two probes show significant cross-hybridization with the same

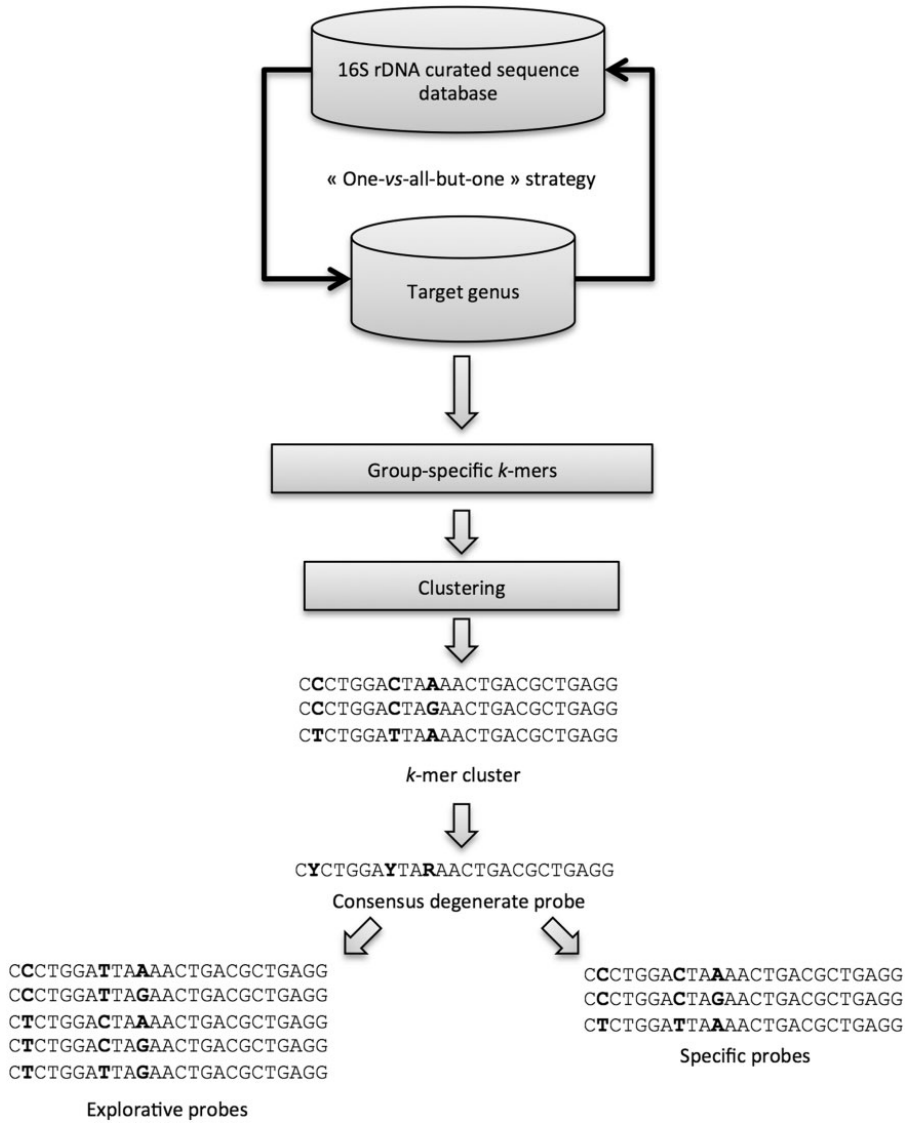


Figure 2. Schematic representation of the KASpOD program workflow.

non-targeted genus, thereby avoiding misleading interpretations of hybridization data.

Finally, after the removal of redundant probes with the PhylGrid probes, 54 129 16S rRNA gene-targeted oligonucleotide probes were added to PhylOPDb.

PhylOPDb web interface

To make all the phylogenetic oligonucleotide probes easily available, we developed a web interface (Figure 3) to fetch and download the 74 003 probes that compose our oligonucleotide database (PhylOPDb). The website, freely available (http://g2im.u-clermont1.fr/phylopdb/), was implemented using PHP, HTML5, CSS3, JavaScript, jQuery, JSON and MySQL. PhylOPDb will be updated annually adding newly designed probes, removing deprecated probes and re-computing coverage and specificity. Moreover, the 16S rDNA sequences databases used for the probe design can be downloaded through the PhylOPDb web interface.

Based on the EMBL taxonomy, the PhylOPDb web interface provides a hierarchical browse of the database contents. When a taxonomic group is selected, corresponding to a kingdom, phylum, class, order, family or genus, the oligonucleotide probes of this group are then displayed. Selected probes can be downloaded both in CSV and FASTA formats. Furthermore, for each non-degenerate probe, associated information is given: identifier, genus

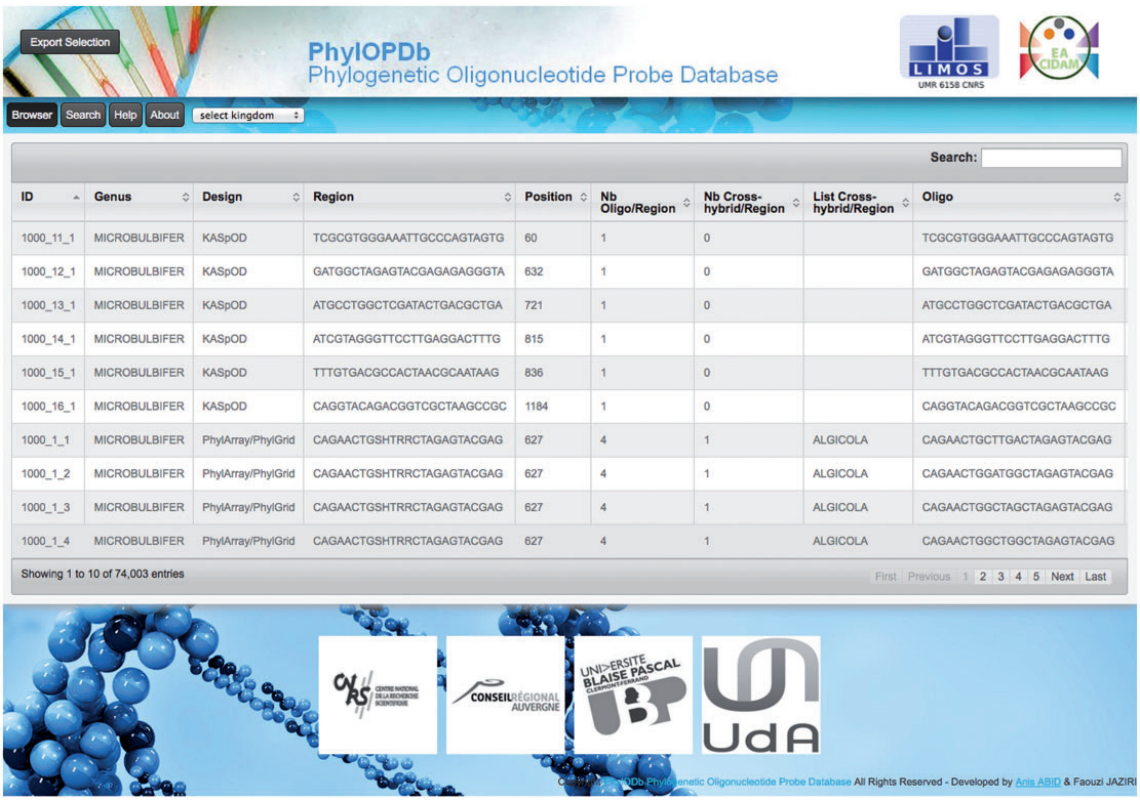


Figure 3. Screenshot of the PhylOPDb web interface.

name, sequence, length, corresponding degenerate probe, degeneracy, coverage percentage, position according to the consensus sequence of the genus, putative cross-hybridizations, melting temperature and the probe design tool used.

Probes can also be obtained using a rapid search by keywords that can represent a part of the sequence of a probe, a genus name or a design name (PhylGrid or KASpOD). Only probes that match these keywords are then displayed. Probes can also be fetched through an advanced search using multiple criteria (genus, design, coverage, melting temperature range or number of cross-hybridizations).

Discussion

Over the past decades, high-throughput molecular tools have opened an unprecedented opportunity for microbiology by enabling the culture-independent genetic study of complex microbial communities, which were so far largely unknown. Among these tools, environmental microarrays, including POAs, are key technologies that are well adapted to profiling environmental communities (21, 24–29). For instance, the PhyloChip (28) is currently the most widely used high-density POA. Nevertheless, 16S rRNA gene-targeted probesets were poorly described and updated. Therefore, PhylOPDb is the most comprehensive SSU rRNA oligonucleotide database by overcoming the currently existing 16S rRNA gene-targeted probe collections. For instance, the entire probeset provided through the PhylOPDb web interface could be used to build a comprehensive POA allowing monitoring of >2000 microbial genera in one experiment. PhylOPDb provides a free and convivial web interface to browse and download a complete 16S rRNA gene-targeted oligonucleotide database composed of 74 003 regular and explorative 25-mers probes covering 2178 prokaryotic genera.

PhylOPDb is also well adapted for other molecular tools using primers or probes (PCR, quantitative PCR, FISH and gene capture) with the availability of group-specific signatures. One of the goals of our database is to exhaustively provide the most specific probes or primers at a fine phylogenetic level. When biologists are interested in specific microbial taxa, it is difficult to reveal them using ‘universal’ probes or primers. For such biological applications, we consider that our database will be helpful. Furthermore, thermodynamics of nucleic acids hybridizations is not fully understood, molecular approaches such as FISH, PCR, DNA microarrays and gene capture are still empiric and need biological validations. With our complete

database, biologists could test various probes or primers to select the most adapted to answer their biological questions. To reduce the complexity of probes and primers selection, we indicate in the database degenerate signatures from which specific probes and primers are deduced. The problem of probes accessibility is particularly relevant for the FISH approach. However, it is possible to resolve this difficulty using helper oligonucleotides, as previously described (30). Recently, a promising strategy for reducing the biocomplexity of environmental samples by enriching the desired genomic target using probes before massive sequencing is being adapted for microbial ecology (3). Most efficient methods rely on the complementary hybridization of oligonucleotide capture probes to the targeted DNA sequences; these methods use capture arrays (31–33), or solution phase, also known as solution hybridization selection (3, 34, 35). Furthermore, the use of explorative probes, as provided by PhylOPDb, in sequence-capture methods allows the full identification and characterization of new taxa (3).

Despite its comprehensiveness, the probeset described in PhylOPDb, however, suffers from a lack of homogeneity that is the third important criterion of a probe design after sensitivity and specificity. Nevertheless, it has been demonstrated that *in silico* approaches for predicting the hybridization behaviour of microarray probes are limited: the only solution is to perform an extensive empirical testing of the probes (36). Moreover, *in silico* assessment of specificity is not sufficient and only an experimental validation can ensure a complete specificity. Prediction of cross-hybridizations strongly relies on both the database and the algorithm. Until now, no clear consensus has been obtained to classify 16S rRNA genes in a unique database. Furthermore, specificity tests suffer from the same complexity without a unique 'universal' tool. Moreover, for some experiments, low melting temperature probes could be used (37). Consequently, we preferred to provide a comprehensive probeset to let the users be able to select the best probes for their own experiments. Therefore, we recommend the PhylOPDb users to combine multiple tools (e.g. BLAST, SILVA-TestProbe, RDP-ProbeMatch) to confirm the specificity results of their selected probes.

Future work will be directed towards the development of specialized sets of phylogenetic oligonucleotides for specific ecosystems. Specificity tests are usually not performed against a suitable subset of sequences, primarily because of the lack of databases for microbial ecology. Depending on the environment studied, it would be more relevant to perform these tests against reduced databanks dedicated to specific ecosystems (e.g. soil, marine, freshwater and gut). Thus, additional probes specific to the targeted environments could be defined. Furthermore, specificity tests will be performed against reduced database limiting computational resources needs.

Probes provided through the PhylOPDb web interface were designed at the genus level but it would be interesting for some purposes such as PCR-based analyses to define probes at different phylogenetic levels. In addition, the current probe collection of the PhylOPDb will be extended to include 18S rRNA gene-targeted oligonucleotide for investigating fungal species within complex environments, where they may play crucial role.

Funding

Auvergne Regional Council, the European Regional Development Fund, the French 'Direction Générale de l'Armement' (DGA) and the program Investissements d'avenir AMI 2011 VALTEX. Funding for open access charge: LIMOS, Clermont Université.

Conflict of interest: None declared.

References

1. Amann,R.I., Ludwig,W. and Schleifer,K.H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59, 143–169.
2. Kirk,J.L., Beaudette,L.A., Hart,M. *et al.* (2004) Methods of studying soil microbial diversity. *J. Microbiol. Methods*, 58, 169–188.
3. Denonfoux,J., Parisot,N., Dugat-Bony,E. *et al.* (2013) Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Res.*, 20, 185–196.
4. DeSantis,T.Z., Hugenholtz,P., Larsen,N. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72, 5069–5072.
5. Quast,C., Pruesse,E., Yilmaz,P. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41, D590–D596.
6. Cole,J.R., Wang,Q., Cardenas,E. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, 37, D141–D145.
7. Ashelford,K.E., Weightman,A.J. and Fry,J.C. (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res.*, 30, 3481–3489.
8. Ludwig,W., Strunk,O., Westram,R. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, 32, 1363–1371.
9. Severgnini,M., Cremonesi,P., Consolandi,C. *et al.* (2009) ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. *Nucleic Acids Res.*, 37, e109.
10. Bader,K.C., Grothoff,C. and Meier,H. (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, 27, 1546–1554.
11. Bader,K.C., Atallah,M.J. and Grothoff,C. (2012) Efficient relaxed search in hierarchically clustered sequence datasets. *J. Exp. Algorithmics (JEA)*, 17, 1.4.

12. Dugat-Bony,E., Peyretaillade,E., Parisot,N. *et al.* (2012) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environ. Microbiol.*, 14, 356–371.
13. Militon,C., Rimour,S., Missaoui,M. *et al.* (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics*, 23, 2550–2557.
14. Missaoui,M., Jaziri,F., Cipiere,S. *et al.* (2011) Large scale parallelization method of 16s rrna probe design algorithm on distributed architecture: application to grid computing. *IEEE*, Bandung (Indonesia). pp. 35–40.
15. Parisot,N., Denonfoux,J., Dugat-Bony,E. *et al.* (2012) KASpOD–a web service for highly specific and explorative oligonucleotide design. *Bioinformatics*, 28, 3161–3162.
16. Jaziri,F., Hill,D.R.C., Parisot,N. *et al.* (2012) MetaExploArrays: a large-scale oligonucleotide probe design software for explorative DNA microarrays. *IEEE Computer Society*, Beijing (China). pp. 664–671.
17. Alm,E.W., Oerther,D.B., Larsen,N. *et al.* (1996) The oligonucleotide probe database. *Appl. Environ. Microbiol.*, 62, 3557–3559.
18. Loy,A., Maixner,F., Wagner,M. *et al.* (2007) probeBase–an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res.*, 35, D800–D804.
19. Altschul,S.F., Gish,W., Miller,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.
20. Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. *et al.* (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, 21, 20–24.
21. Loy,A., Lehner,A., Lee,N. *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl. Environ. Microbiol.*, 68, 5064–5081.
22. McDonald,D., Price,M.N., Goodrich,J.K. *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, 6, 610–618.
23. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659.
24. Wilson,K.H., Wilson,W.J., Radosevich,J.L. *et al.* (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl. Environ. Microbiol.*, 68, 2535–2541.
25. Brodie,E.L., DeSantis,T.Z., Joyner,D.C. *et al.* (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl. Environ. Microbiol.*, 72, 6288–6298.
26. Palmer,C., Bik,E.M., Eisen,M.B. *et al.* (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Res.*, 34, e5.
27. Brodie,E.L., DeSantis,T.Z., Parker,J.P.M. *et al.* (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc. Natl Acad. Sci. USA*, 104, 299–304.
28. DeSantis,T.Z., Brodie,E.L., Moberg,J.P. *et al.* (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb. Ecol.*, 53, 371–383.
29. Hazen,T.C., Dubinsky,E.A., DeSantis,T.Z. *et al.* (2010) Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science*, 330, 204–208.
30. Fuchs,B.M.,Glöckner,F.O.,Wulf,J. *et al.* (2000) Unlabeled helper oligonucleotides increase the *in situ* accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, 66, 3603–3607.
31. Albert,T.J., Molla,M.N., Muzny,D.M. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, 4, 903–905.
32. Okou,D.T., Steinberg,K.M., Middle,C. *et al.* (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, 4, 907–909.
33. Mokry,M., Feitsma,H., Nijman,I.J. *et al.* (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res.*, 38, e116.
34. Gnirke,A., Melnikov,A., Maguire,J. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, 27, 182–189.
35. Tewhey,R., Nakano,M., Wang,X. *et al.* (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, 10, R116.
36. Loy,A. and Bodrossy,L. (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin. Chim. Acta*, 363, 106–119.
37. Kaplinski,L., Scheler,O., Parkel,S. *et al.* (2010) Detection of tmRNA molecules on microarrays at low temperatures using helper oligonucleotides. *BMC Biotechnol.*, 10, 34.

3.4 Discussion

PhylOPDb propose aujourd'hui la collection de sondes oligonucléotidiques, ciblant l'ADNr 16S, la plus exhaustive. Composée de 74 003 sondes sensibles, spécifiques et exploratoires, PhylOPDb est librement consultable *via* une interface web conviviale. Les sondes répertoriées peuvent être utilisées pour l'identification taxonomique de plusieurs milliers de genres bactériens au travers de différentes méthodes moléculaires comme la PCR, le FISH, la capture de gènes ou les biopuces ADN. En effet, lorsque l'on s'intéresse à un groupe taxonomique en particulier, il est souvent très contraignant de définir ses propres amorces et encore moins pertinent d'utiliser des sondes ou amorces dites universelles. Grâce à PhylOPDb, il est désormais possible pour les biologistes de pouvoir disposer de sondes définies à un niveau taxonomique aussi précis que celui du genre.

Cependant, la détermination *in silico* des critères essentiels aux sondes oligonucléotidiques comme la sensibilité, la spécificité ou l'uniformité reste délicate sans validation biologique (Loy & Bodrossy 2006). Les utilisateurs doivent donc rester prudents, en particulier pour ce qui concerne la prédiction d'hybridations croisées potentielles.

De nombreuses évolutions sont envisageables pour PhylOPDb avec notamment l'ajout de nouvelles sondes définies à d'autres niveaux taxonomiques ou à partir d'autres biomarqueurs phylogénétiques comme l'ADNr 18S pour l'étude des communautés eucaryotes. Par ailleurs, pour permettre une meilleure discrimination taxonomique des espèces il est parfois nécessaire d'avoir recours à l'analyse de plusieurs biomarqueurs. Ces stratégies, dites MLSA (*Multi-Locus Sequence Analysis*), se basent généralement sur des gènes fonctionnels comme les gènes dit de ménage. La détermination de sondes ciblant des gènes codants pour des protéines doit donc également faire l'objet de développements bioinformatiques importants en écologie microbienne.

4. Développement d'un logiciel de détermination de sondes oligonucléotidiques ciblant des gènes fonctionnels

4.1 Contexte

L'étude des capacités métaboliques des populations microbiennes présentes dans l'environnement revêt aujourd'hui de nombreux intérêts non seulement au niveau fondamental pour la compréhension du fonctionnement des écosystèmes mais aussi au niveau économique et industriel (Steele *et al.* 2009). Les biopuces ADN fonctionnelles (FGA) sont particulièrement pertinentes pour répondre à cette problématique puisque leur hybridation, à la fois avec des cibles ADN et ARN extraites des mêmes échantillons environnementaux, permet de percevoir très rapidement quels sont les gènes exprimés parmi le répertoire génique de la microflore présente. Cependant, pour que ces approches présentent un avantage indéniable, il faut qu'elles puissent donner accès à la fraction inconnue des populations microbiennes que l'on sait aujourd'hui considérable. Ceci est dorénavant possible grâce au développement de nouveaux logiciels qui autorisent la détermination de sondes dites exploratoires. Avant nos travaux, les deux seuls logiciels permettant de déterminer des sondes exploratoires dédiées à l'élaboration de FGA décrits dans la littérature étaient Metabolic Design (Terrat *et al.* 2010) et HiSpOD (Dugat-Bony *et al.* 2011).

La stratégie Metabolic Design est basée sur l'utilisation d'alignements de séquences protéiques pour l'identification de motifs conservés. Ces motifs d'acides aminés sont alors rétro-traduits en nucléotides pour la sélection de sondes en intégrant toute la dégénérescence du code génétique. La difficulté de cette approche est la capacité à déterminer des sondes longues (30-mers et plus) pour construire des FGA très sensibles même pour des cibles peu abondantes, ce qui est généralement le cas pour des gènes fonctionnels et leurs transcrits (Gentry *et al.* 2006). La traduction inverse des motifs protéiques supérieurs à 7 ou 8 acides aminés génère le plus souvent des séquences nucléiques avec des taux de dégénérescence très élevés. Le nombre de sondes non dégénérées issues de ces séquences peut alors être considérable, et encore inadapté aux formats de biopuces disponibles actuellement sur le marché. En effet, si l'on considère que chaque acide aminé est codé en moyenne par trois codons différents (20 acides aminés spécifiés par 61 codons), chaque sonde dégénérée de 51 mers (soit l'équivalent de 17 acides aminés) produit plus de 3^{17} combinaisons non dégénérées (soit plus de 129 millions).

La stratégie HiSpOD est basée sur la sélection de sondes à partir i) de séquences nucléiques correspondant à des gènes, ou ii) de séquences consensus dégénérées, issues de l'alignement de toutes les séquences nucléiques disponibles pour une famille de gènes. Comparée à l'approche utilisant des séquences protéiques en entrée, cette stratégie génère des sondes beaucoup moins dégénérées ce qui autorise la sélection de sondes de taille supérieure.

Cependant, ces stratégies ne permettent pas de gérer les grands jeux de données produits par la généralisation des approches haut-débit en écologie microbienne. En effet, l'étape d'alignement multiple permettant d'identifier des régions conservées devient rapidement irréalisable, aussi bien en terme de temps de calcul qu'en terme de qualité d'alignement. Afin de s'affranchir de cette contrainte, de nouvelles stratégies basées sur l'identification de motifs courts (*k*-mers) spécifiques des séquences ciblées tendent à être développées (Hysom *et al.* 2012 ; Parisot *et al.* 2012).

4.2 Objectif

Afin d'apporter une alternative innovante à la conception de sondes pour biopuces fonctionnelles, l'objectif de ce travail de recherche était de développer un nouveau logiciel alliant les avantages des deux méthodes précédemment citées tout en étant capable d'intégrer la masse de données génomiques actuelle. En effet, la manipulation et l'organisation des données nécessaires à l'obtention de sondes oligonucléotidiques spécifiques de gènes d'intérêt ne sont pas toujours des tâches aisées pour le biologiste puisqu'elles impliquent l'extraction pertinente des données et la manipulation de plusieurs outils bioinformatiques. L'outil développé doit donc être capable de récupérer et d'organiser automatiquement ces données de séquences avant de procéder à la détermination de sondes.

La stratégie employée permet ainsi, à partir d'une recherche par mots-clés, de récupérer l'ensemble des séquences nucléiques et protéiques correspondant au gène ciblé, d'organiser ces données afin de constituer des groupes taxonomiques et fonctionnels, de rechercher des motifs protéiques conservées au sein des différents groupes et d'en déduire des sondes oligonucléotidiques dégénérées. Basée sur l'utilisation de *k*-mers et intégrant le biais d'usage du code génétique des groupes taxonomiques constitués pour limiter la dégénérescence, cette approche permet donc de générer des sondes oligonucléotidiques longues, à partir de motifs protéiques, en s'affranchissant de toute étape d'alignement multiple. L'ensemble de la démarche mise en place a conduit à l'élaboration d'un logiciel

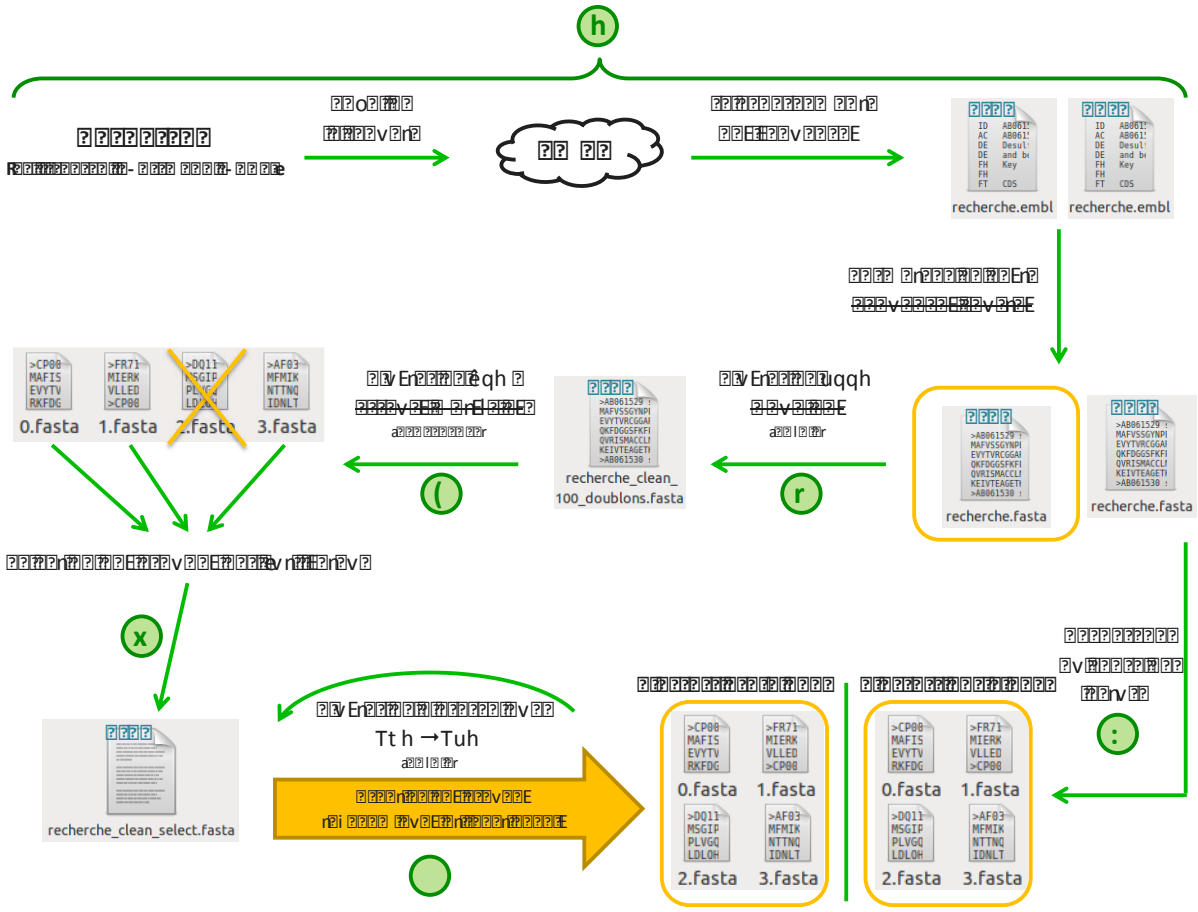


Figure 10. Schéma récapitulatif de la fouille de données et de la construction des *clusters* au sein du logiciel ProKSpOD.

L'étape **(1)** représente la recherche par mots-clés au sein de la base de données EMBL et le rapatriement des séquences correspondantes. Les étapes **(2)**, **(3)** et **(4)** consistent en l'élimination des doublons et des erreurs d'affiliation. L'étape **(5)** permet la création de groupes taxonomiques et fonctionnels au travers d'un *clustering* hiérarchique des séquences. Enfin, l'étape **(6)** contrôle le cadre de lecture de chacune des séquences restantes.

nommé ProKSpOD pour : « *Protein-coding sequence based K-mer algorithm for high-Specific Oligonucleotide Design* ».

4.3 Principaux résultats

Le travail réalisé au cours de cette étude a permis de proposer un nouveau logiciel de détermination de sondes oligonucléotidiques dédié aux gènes fonctionnels. Cette stratégie, toujours en cours de développement, peut se détailler de la façon suivante.

4.3.1 Fouille de données et construction des *clusters*

La première étape de l'algorithme consiste à récupérer, de manière automatique, l'ensemble des séquences nucléiques et protéiques correspondant au gène d'intérêt. Pour cela, le logiciel RAA_QUERY (Gouy & Delmotte 2008) est utilisé afin d'effectuer une recherche par mots-clés au sein de la base de données de l'EMBL (Brooksbank *et al.* 2014) (**Figure 10, point 1**).

Cependant, l'un des inconvénients majeurs des bases de données internationales réside dans la redondance des informations. Ainsi, la deuxième étape de la stratégie consiste à éliminer l'ensemble des séquences strictement redondantes (*i.e.* séquences identiques ou séquences courtes totalement incluses dans d'autres séquences plus longues) par l'intermédiaire d'un *clustering* à 100% d'identité protéique, *via* l'outil CD-HIT (Fu *et al.* 2012), des séquences récupérées précédemment (**Figure 10, point 2**). Cette étape permet donc de réduire la complexité de l'analyse en diminuant le nombre de séquences. De plus, les séquences non informatives car trop courtes (moins de 60 acides aminés) sont également éliminées de l'analyse.

La qualité d'annotation fonctionnelle des séquences est une autre limite importante de ces bases de données. En effet, il se peut que la recherche par mots-clés ait permis la récupération de séquences ne correspondant pas biologiquement au gène étudié, mais annotées comme telles. L'étape suivante consiste donc à identifier ces séquences mal annotées, qui sont donc éloignées du gène d'intérêt du point de vue de leurs séquences. Pour cela, un nouveau *clustering* est réalisé sur les séquences protéiques grâce à l'outil BLASTCLUST (Altschul *et al.* 1990) à un seuil faible de 30% d'identité protéique (**Figure 10, point 3**). Dans cette situation, l'utilisation du logiciel BLASTCLUST et non CD-HIT est nécessaire car ce dernier ne permet pas de réaliser des *clustering* à des seuils d'identité protéique aussi faibles.

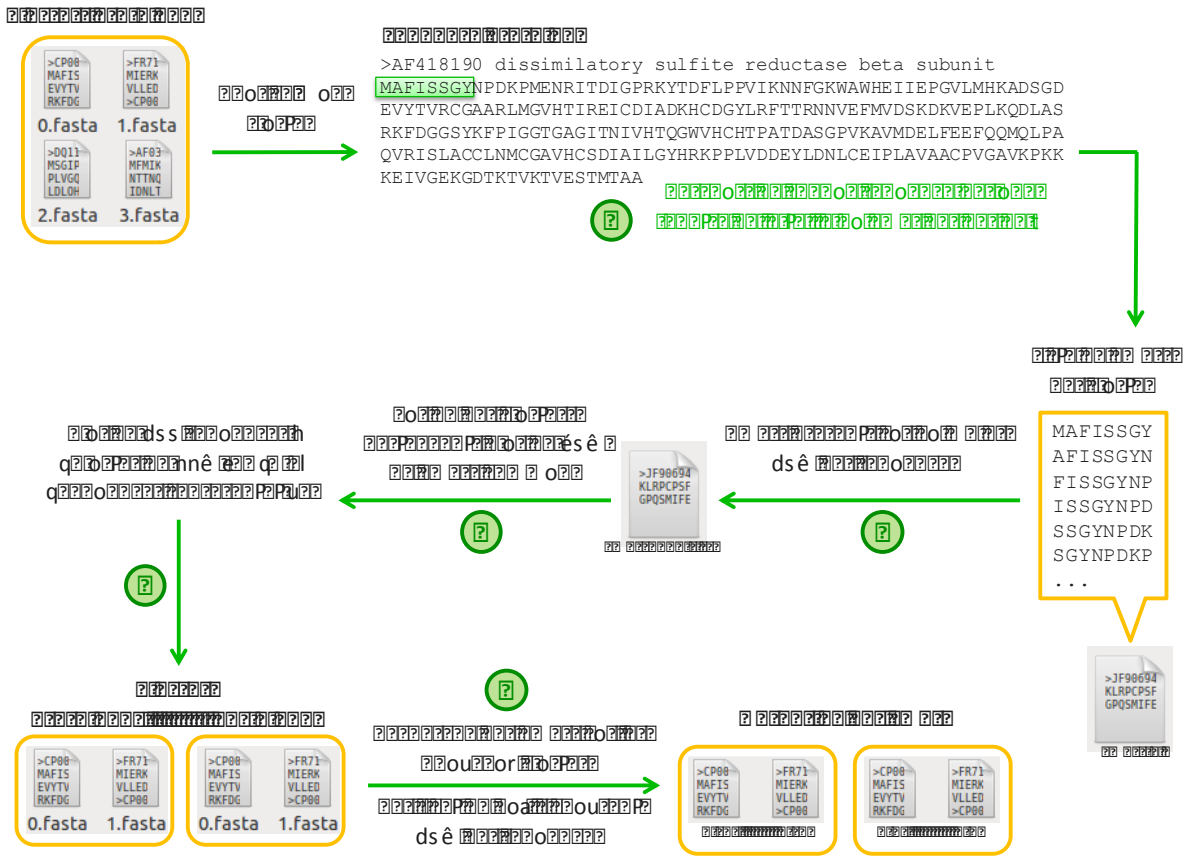


Figure 11. Schéma récapitulatif de la recherche des *k*-mers au sein du logiciel ProKSpOD.

L'étape **(1)** consiste en l'extraction de tous les *k*-mers par *cluster*. A l'issue de l'étape **(2)** seuls les *k*-mers présents dans au moins 50% des séquences sont conservés. L'étape **(3)** fusionne les clusters partageant au moins 70% de *k*-mers puis l'étape **(4)** se charge de diminuer la complexité du traitement en réduisant la taille des *clusters* trop peuplés. Enfin l'étape **(5)** extrait tous les *k*-mers dans les *clusters* nouvellement construits.

A la fin de cette étape, une intervention de l'utilisateur est nécessaire pour sélectionner les *clusters* qui correspondent effectivement à la recherche et s'affranchir de ceux qui résultent d'erreurs d'annotation (**Figure 10, point 4**).

Afin d'obtenir une biopuce fonctionnelle la plus pertinente possible, il est important que le *design* de sondes puisse détecter des gènes présents dans l'environnement, mais également les groupes microbiens portants ces gènes. Ainsi, il est nécessaire de regrouper les séquences taxonomiquement proches. Pour cela, un *clustering* hiérarchique (CD-HIT) est appliqué, afin de regrouper les séquences proches tout en minimisant le nombre de *clusters* ne comportant qu'une seule séquence, ou un nombre trop important de séquences. Pour ce faire, les séquences sont, tout d'abord, regroupées à un seuil élevé d'identité protéique (par défaut 97%). Les *clusters* comportant plusieurs séquences sont alors conservés, tandis que les séquences seules sont remises en commun et subissent un nouveau *clustering* à un seuil plus faible. Ces mêmes étapes sont ensuite répétées avec une diminution du seuil, jusqu'à l'atteinte du seuil minimal fixé à 91% (**Figure 10, point 5**).

Finalement, avant de procéder à la détermination des sondes, il est nécessaire de définir le bon cadre de lecture pour les séquences nucléiques. Pour cela, au sein de chaque *cluster*, toutes les séquences nucléiques sont comparées à une séquence protéique représentative du *cluster* grâce à une analyse BLASTX. Les séquences nucléiques peuvent alors être tronquées pour que le premier cadre de lecture corresponde au CDS ciblé (**Figure 10, point 6**).

4.3.2 Recherche des *k*-mers

Cette étape est réalisée à partir des séquences protéiques de chaque *cluster*. Pour cela, chaque séquence est parcourue par une fenêtre de taille *k* avec un pas de un acide aminé (**Figure 11, point 1**). Ceci nous permet d'obtenir une liste de *k*-mers par *cluster*, parmi laquelle les *k*-mers présents dans moins de 50% des séquences seront supprimés (**Figure 11, point 2**). Tous les *clusters* partageant au moins 70% de *k*-mers sont également fusionnés (**Figure 11, point 3**). En effet, les *clusters* possédant un trop grand nombre de *k*-mers en commun seront difficilement discriminables par les sondes oligonucléotidiques. Après cette étape, les *clusters* contenant plus de 500 séquences subissent un nouveau *clustering* (*i.e.* CD-HIT à 99% d'identité protéique) afin de n'en conserver que les séquences les plus représentatives, dans le but de réduire le nombre d'informations à traiter dans la suite de l'algorithme (**Figure 11, point 4**). Enfin, des *k*-mers les plus représentatifs des *clusters*

Tableau 4. Seuils limites d'utilisation des codons choisis pour l'algorithme ProKSpOD. En fonction du nombre de possibilités par acide aminé, un seuil minimal d'utilisation des codons est fixé. Tout codon utilisé à une fréquence inférieure à ce seuil sera considéré comme rare pour ce *cluster*, et sera éliminé du code génétique favori

Nombre de codons possibles	Seuil limite d'utilisation des codons
1	100%
2	30%
3	
4	10%
6	

nouvellement construits sont de nouveau sélectionnés et sont conservés pour la rétrotraduction (**Figure 11, point 5**).

4.3.3 Rétrotraduction des *k*-mers

L'approche classique consisterait alors à effectuer la traduction inverse des *k*-mers protéiques sélectionnés en intégrant toute la dégénérescence du code génétique. Cependant, à partir d'une sonde dégénérée, chaque combinaison spécifique est fixée sur la biopuce. Les formats de biopuces étant actuellement limités à un million de sondes, il est alors nécessaire de déterminer le jeu de sondes le plus réduit possible mais permettant de cibler toute la diversité de séquences des gènes ciblés. Ainsi, afin de limiter la dégénérescence des sondes déterminées, la stratégie mise au point repose sur l'évaluation du biais d'usage des codons synonymes au sein des séquences de chaque *cluster*.

Pour cela, une fréquence d'utilisation de chaque codon est calculée à partir de toutes les séquences nucléiques du *cluster*. Suivant le nombre de codons possibles pour un acide aminé donné, un pourcentage d'utilisation limite est fixé pour éliminer les codons faiblement représentés (**Tableau 4**). Ainsi, on obtient un code génétique que l'on peut qualifier de « favori » pour le *cluster* traité. Cependant, comme un codon considéré comme rare (*i.e.* fréquence d'utilisation inférieure au seuil fixé) peut être utilisé pour coder un acide aminé au sein du motif protéique considéré, une vérification doit être réalisée. Si tel est le cas, ce codon rare sera temporairement ajouté au code génétique « favori » pour réaliser la traduction inverse du *k*-mer. Enfin, chaque position de la séquence nucléique est alors évaluée selon le standard IUB/IUPAC (Cornish-Bowden 1985) et la sonde dégénérée est créée.

Pour limiter encore la dégénérescence des sondes, le dernier nucléotide de la séquence, correspondant à la troisième base d'un codon (*i.e.* base la plus variable des codons synonymes), est éliminé. La dégénérescence est calculée de manière multiplicative et seules les sondes ayant une dégénérescence acceptable (inférieure à 128) sont conservées.

La partie comprenant la recherche des *k*-mers et leur rétrotraduction est une étape itérative de l'algorithme. Dans un premier temps, la recherche des *k*-mers est effectuée en se basant sur une taille maximale de *k*-mer de 17 acides aminés. Si cette taille ne permet pas d'aboutir à une détermination de sondes suffisante pour tous les *clusters* (au moins 5 sondes par *cluster*), les différentes étapes sont répétées en réduisant avec un pas de 1 la taille des *k*-

mers. L'itération s'arrête lorsque tous les *clusters* sont suffisamment couverts, ou si la taille minimale de *k*-mer (*i.e.* 8 acides aminés) est atteinte.

4.3.4 Vérification des sondes obtenues

Une fois les sondes définies pour chaque *cluster*, celles-ci doivent être vérifiées en termes de couverture, spécificité et chevauchement.

Pour réaliser ce test, les sondes déterminées sont alignées contre l'ensemble des séquences nucléiques du *cluster* grâce à l'outil PatMaN (Prüfer *et al.* 2008). Un pourcentage de couverture est défini en rapportant le nombre de séquences uniques couvertes par la sonde au nombre total de séquences du *cluster*. Les sondes ayant le taux de couverture le plus faible (*i.e.* dernier quartile) ne seront alors pas conservées.

Le test de spécificité permet de vérifier que les sondes sélectionnées n'entraineront pas d'appariements non désirés lors de l'hybridation. Afin de réaliser ce test de spécificité, la base de données EnvExBase (Dugat-Bony *et al.* 2011) dédiée à l'écologie microbienne a été utilisée. Il s'agit d'une base de données de CDS complète élaborée à partir des séquences comprises dans les divisions procaryote (PRO), champignon (FUN) et environnement (ENV) de la base de données de séquences nucléiques EMBL. Les sondes sont alors comparées, grâce à l'outil PatMaN, à l'intégralité des séquences de cette base, et la possibilité d'hybridations croisées potentielles peut être vérifiée.

Enfin, un test de chevauchement est réalisé afin de vérifier que chaque sonde oligonucléotidique possède au moins 12 bases uniques non chevauchantes avec les autres sondes.

4.4 Discussion

Le logiciel ProKSpOD offre donc de nouvelles perspectives en écologie microbienne en comparaison aux autres outils de détermination de sondes pour biopuces fonctionnelles (*i.e.* Metabolic Design (Terrat *et al.* 2010) et HiSpOD (Dugat-Bony *et al.* 2011)).

Grâce à cette stratégie, il est possible d'automatiser toutes les étapes préalables à la détermination qui peuvent s'avérer fastidieuses. L'utilisateur précise simplement le nom du gène sur lequel il souhaite travailler et l'algorithme se charge de récupérer et d'organiser automatiquement les données de séquences associées. A partir de ces informations de séquences, et peu importe la masse de données qu'elles représentent, le logiciel ProkSpOD est

capable d'identifier des motifs protéiques conservés pour assurer une détermination de sondes oligonucléotidiques. Cependant, l'approche développée au travers de ce logiciel est différente de celle proposée par Metabolic Design, pour lequel les sondes, déterminées après traduction inverse de motifs protéiques, sont généralement courtes pour éviter un niveau de dégénérescence trop élevé. ProKSpOD tire profit du biais d'usage du code génétique qu'il existe chez toutes les espèces microbiennes pour sélectionner des sondes moins dégénérées..

La possibilité de tester la spécificité de chaque sonde contre une base de données de CDS spécialisée (*i.e.* plus de 10 millions de séquences), procure au logiciel ProKSpOD, comme pour HiSpOD, un avantage supplémentaire non négligeable pour la sélection de sondes très spécifiques. Cependant, ce test de spécificité est l'étape qui impacte le plus sur les temps de calculs nécessaires pour le *design*. Actuellement installé sur une machine multiprocesseurs de 135 CPUs, il est possible d'envisager de subdiviser la base de données suivant le type d'environnement étudié (sol, eau douce, mer, air etc.), ce qui devrait conduire à la réduction des temps de calculs et ainsi permettre la détermination de sondes pour un très grand nombre de gènes.

Comme évoqué précédemment, la qualité d'annotation fonctionnelle fait parfois défaut aux bases de données de séquences, et la recherche par mots-clés peut donc entrainer l'omission de nombreuses séquences. Cependant, il serait intéressant de pouvoir prendre en compte l'ensemble de ces séquences des bases de données n'ayant aucune annotation fonctionnelle ou ayant été mal annotées, mais correspondant néanmoins au gène étudié. Pour cela, il serait possible d'utiliser chacune des séquences protéiques des groupes taxonomiques formés afin de réaliser une fouille de données exhaustive par exemple contre la base de données de CDS de l'EMBL.

L'ensemble des logiciels de détermination de sondes oligonucléotidiques développés au cours de cette thèse répond aux problématiques posées par l'écologie microbienne à l'heure actuelle. En étant capables de gérer d'importantes masses de données pour permettre la sélection de sondes à la fois sensibles, spécifiques et exploratoires, ces stratégies de *design* peuvent être appliquées à l'étude phylogénétique et fonctionnelle d'environnements complexes au travers d'outils moléculaires ou bioinformatiques variés.

PARTIE 3 : Applications moléculaires et bioinformatiques

1. Développement de biopuces phylogénétiques environnementales

1.1 Contexte

Comme nous l'avons vu précédemment, les techniques moléculaires classiquement utilisées pour étudier les communautés microbiennes peuvent s'avérer inadaptées pour analyser globalement la diversité bactérienne d'environnements complexes. Une technique alternative repose sur l'utilisation de biopuces ADN pour permettre l'analyse simultanée de millions de gènes en une seule expérience.

Les biopuces ADN ont été utilisées dans de nombreux domaines des sciences de la vie, y compris en écologie microbienne (Zhou 2003 ; Gentry *et al.* 2006 ; Wagner *et al.* 2007 ; Roh *et al.* 2010 ; Parisot *et al.* 2014). Il en existe différentes catégories adaptées aux nombreuses problématiques de génomique environnementale (Gentry *et al.* 2006). Ainsi, la première catégorie de biopuces, appelée « *Whole Genome Array* » (WGA), permet de cibler l'ensemble des gènes d'un ou plusieurs microorganismes et peut être utilisée pour caractériser des souches ou des *consortia* isolés de l'environnement (*e.g.* l'étude de 4 génomes de microorganismes anaérobies en réponse à un stress oxydatif (Scholten *et al.* 2007)). Cependant l'utilisation de ces biopuces pour l'étude *in situ* d'échantillons environnementaux est généralement limitée du fait de la complexité des communautés microbiennes composées majoritairement de souches pour lesquelles il n'existe pas de données de séquences sur leur génome. Il est alors plus judicieux d'utiliser des biopuces environnementales comme les biopuces phylogénétiques (*Phylogenetic Oligonucleotide Array* ou POA) ou les biopuces fonctionnelles (*Functional Gene Array* ou FGA).

Afin de rapidement caractériser la structure des communautés microbiennes au sein d'environnements complexes, de nombreuses biopuces phylogénétiques ciblant le gène biomarqueur ADNr 16S ont été développées. On peut alors distinguer des POAs généralistes comme la PhyloChip (Brodie *et al.* 2006), qui couvre la quasi-totalité de la diversité procaryotique répertoriée dans les bases de données, et des POAs spécifiques à l'étude d'un groupe taxonomique ou d'un environnement donné. C'est par exemple le cas de la HITChip

(Rajilić-Stojanović *et al.* 2009 ; 2012) qui permet l'étude phylogénétique du microbiote intestinal humain.

1.2 Objectif

En tirant profit des algorithmes de détermination de sondes oligonucléotidiques développés au cours de cette thèse, l'objectif de ce travail a consisté à mettre en place une approche biopuce phylogénétique pour l'étude des communautés bactériennes d'environnements complexes.

Dans un premier temps, une biopuce phylogénétique dédiée à l'étude du microbiote intestinal humain a été développée. Cet environnement, le plus abondant et diversifié du corps humain, a fait l'objet de plusieurs études *via* l'utilisation de biopuces phylogénétiques (Paliy & Agans 2012). Néanmoins, toutes ces études ne s'intéressent qu'aux microorganismes déjà caractérisés dans les bases de données. Afin de permettre une étude exhaustive des communautés bactériennes qui composent cet environnement, une détermination de sondes exploratoires a été effectuée à l'aide de l'outil PhylArray (Militon *et al.* 2007). L'ensemble de ces sondes a permis le développement de la biopuce nommée HuGChip pour « *Human Gut Chip* ».

En parallèle, une biopuce phylogénétique généraliste intégrant les deux jeux de sondes obtenus grâce aux algorithmes PhylGrid 2.0 et KASpOD, décrits au sein de la PARTIE 2 : Détermination de sondes oligonucléotidiques, a été mise au point. Il s'agit de la première biopuce phylogénétique exhaustive possédant un caractère exploratoire pouvant ainsi permettre l'étude des microorganismes, connus ou non, au sein de divers environnements.

1.3 Principaux résultats

1.3.1 Biopuce HuGChip

En utilisant la base de données de 1052 séquences d'ADNr 16S issues du microbiote intestinal humain, développée par (Rajilić-Stojanović *et al.* 2009), 4441 sondes de 25 nucléotides ont pu être obtenues afin de construire la biopuce HuGChip. Cette biopuce permet ainsi l'étude simultanée de 66 familles bactériennes représentatives du microbiote intestinal humain. Pour chacune de ces familles, 5 régions différentes du gène ADNr 16S ont été ciblées grâce au logiciel PhylArray (Militon *et al.* 2007). Ce niveau taxonomique a été choisi afin de garantir le multiplexage maximal (*i.e.* jusqu'à 16 échantillons traités en parallèle) tout en

conservant une précision d'analyse intéressante (*i.e.* information phylogénétique). La plateforme *Agilent Technologies* 8×15k a donc été choisie pour cette biopuce afin de permettre une synthèse des sondes en triplicat pour assurer une interprétation robuste des résultats.

La validation biologique de cette biopuce phylogénétique a été effectuée en deux temps. Tout d'abord, afin de déterminer les seuils optimaux à appliquer pour l'analyse des résultats de la HuGChip, un mélange connu de produits d'amplification PCR du gène ADNr 16S provenant de 5 souches bactériennes a été hybridé. Après analyse, seuls les signaux supérieurs à 12 fois le bruit de fond ont été considérés et un seuil minimal de détection de 3 sondes sur les 5 déterminées par famille bactérienne a été choisi. Ces mêmes seuils ont ensuite été appliqués pour l'étude d'échantillons de selles humaines et les résultats ont pu être comparés avec une approche par séquençage haut-débit (*i.e.* métagénomique et amplicons dirigés sur le gène ADNr 16S). L'analyse montre une corrélation importante des résultats et certains taxa uniquement détectés par la HuGChip ont pu être validés par PCR quantitative démontrant la pertinence de l'approche. Le travail réalisé a donné lieu à une publication dans le journal « *PLoS One* ».

Article n°6

The Human Gut Chip 'HuGChip', an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level.

OPEN ACCESS Freely available online

PLOS ONE

The Human Gut Chip "HuGChip", an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level

William Tottey[1], Jeremie Denonfoux[1], Faouzi Jaziri[1,2], Nicolas Parisot[1], Mohiedine Missaoui[1,2], David Hill[2], Guillaume Borrel[1], Eric Peyretaillade[1], Monique Alric[1], Hugh M. B. Harris[3], Ian B. Jeffery[3], Marcus J. Claesson[3], Paul W. O'Toole[3], Pierre Peyret[1], Jean-François Brugère[1]*

1 EA CIDAM 4678, Clermont-Université, Université d'Auvergne, Clermont-Ferrand, France, **2** CNRS, UMR 6158, ISIMA/LIMOS, Aubière/Clermont-Ferrand, France, **3** Department of Microbiology and Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland

Abstract

Evaluating the composition of the human gut microbiota greatly facilitates studies on its role in human pathophysiology, and is heavily reliant on culture-independent molecular methods. A microarray designated the Human Gut Chip (HuGChip) was developed to analyze and compare human gut microbiota samples. The PhylArray software was used to design specific and sensitive probes. The DNA chip was composed of 4,441 probes (2,442 specific and 1,919 explorative probes) targeting 66 bacterial families. A mock community composed of 16S rRNA gene sequences from intestinal species was used to define the threshold criteria to be used to analyze complex samples. This was then experimentally verified with three human faecal samples and results were compared (i) with pyrosequencing of the V4 hypervariable region of the 16S rRNA gene, (ii) metagenomic data, and (iii) qPCR analysis of three phyla. When compared at both the phylum and the family level, high Pearson's correlation coefficients were obtained between data from all methods. The HuGChip development and validation showed that it is not only able to assess the known human gut microbiota but could also detect unknown species with the explorative probes to reveal the large number of bacterial sequences not yet described in the human gut microbiota, overcoming the main inconvenience encountered when developing microarrays.

Citation: Tottey W, Denonfoux J, Jaziri F, Parisot N, Missaoui M, et al. (2013) The Human Gut Chip "HuGChip", an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level. PLoS ONE 8(5): e62544. doi:10.1371/journal.pone.0062544

Editor: Mark R. Liles, Auburn University, United States of America

Received January 15, 2013; **Accepted** March 22, 2013; **Published** May 17, 2013

Funding: This work was supported by a PhD studentship supported by the European Union and the Auvergne Council to WT. MJC, HH, IBG, and PWOT are members of the ELDERMET consortium (http://eldermet.ucc.ie) whose work is supported in part by the (Govt. of Ireland) Dept. Agriculture, Fisheries, and Food/ Health Research Board FHRI award to the ELDERMET project, as well as the Alimentary Pharmabiotic Center. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jf.brugere@udamail.fr

Introduction

The human gut harbours a complex ecosystem composed of 10^{14} microbial cells [1], including eukaryotic and archaeal cells [2,3]. Although a high inter-individual diversity is present and is modulated by several factors [4–6], a phylogenetic core at the species level was hypothesized [7]:composed of 66 Operational Taxonomic Units (OTUs) which were present in more than 50% of the individuals and which represented about 36% of the total sequences. More than 1,500 different bacterial species have already been associated with the human gut microbiota and around 500 different bacterial species constitute an individual human gut microbiota [8]. Furthermore, it has been shown that the gut microbiota impacts upon the health of its host, for example by influencing the maturation of the immune system, by modulating the barrier function the gut epithelium and by conferring colonization resistance or direct antagonism protection against pathogens [9]. It also provides a set of metabolic functions which are not present in the coding capacity of human organism, such as the digestion of some resistant carbohydrates, energy storage or the production of vitamins [10]. Furthermore, the gut microbiota has also been reported to play a major role in diseases like colon cancer [11], obesity [12], inflammatory bowel disease [13,14] or cardiovascular disease [15]. Over the last two decades, development of culture independent techniques has significantly increased our knowledge of gut microbiota. Tools permitting exhaustive analysis of individual gut microbiota including a phylogenetic identification and (semi-) quantification are still under development. Most of these techniques are based on the 16S ribosomal RNA (rRNA) gene sequence variations between different species. Fluorescence In Situ Hybridization (FISH) and fingerprinting techniques such as Denaturing Gradient Gel Electrophoresis (DGGE), Terminal Fragment Length Polymorphism (T-RFLP) are frequently used (reviewed in [16]). However, they generally lack resolution and do not allow high-throughput direct phylogenetic identification. More recently techniques such as DNA microarray hybridization and next-generation sequencing (NGS) have been developed granting further phylogenetic identification of microbiota diversity [16,17].

Microarray technology is a high throughput platform used to study numerous samples and to detect thousands of nucleic acids sequences simultaneously making it fast and user friendly.

Phylogenetic DNA microarrays consist of several thousand probes, usually designed from rRNA gene sequence database targeting either specific organisms (e.g. pathogenic bacteria) or the whole microbiota at various taxonomic levels. The use of 16S rRNA microarrays provides superior diagnostic power compared to clone library techniques [18]. Several microarrays addressing the gut microbiota have been developed over the last decade, showing differences in their design and the aims of study. In 2007, Palmer and colleagues designed an array containing 10,265 probes, each spotted once, and targeting 1,629 species [19]. Another microarray addressing the whole gut microbiota was published by Paliy *et al.* (2009) and was spotted with 16,223 probes targeting 775 bacterial species [20]. Finally, the Human Intestinal Tract Chip (HITChip) was designed to target 1,140 species using 4,809 overlapping probes [21]. More recently, array hybridization results were compared to pyrosequencing of the V1 to V6 hypervariable regions of the 16S rRNA gene sequence and showed a good correlation [22,23]. The authors suggested that the differences observed between the data from the two techniques might arise from a combination of the analysis of different hypervariable regions, the limited number of 16S rRNA gene sequences available for the probe design, and the ability of these probes to only target known 16S rRNA gene sequences.

Phylogenetic microarray probe design can be performed using various software packages such as ARB [24], PRIMROSE [25] and ORMA [26] which have been widely used as they provide specific and sensitive probes to address sequences from databases. In spite of the exponential growth of data within international databases, our current understanding of microbial diversity is still incomplete. Microarrays coupled with explorative probe design strategies are, therefore, well suited to survey complete microbial communities, including microorganisms with uncharacterized sequences [27]. The PhylArray [28] and the KASpOD [29] probe design software were developed to provide sensitive, specific and also explorative probes dedicated to phylogenetic microarrays [28]. This innovative probe design strategy may help to overcome the main limitation of microarrays i.e. the inability to detect unknown sequences and thus, to survey uncharacterized microbial populations.

In this study, we present the Human Gut Chip (abbreviated in HuGChip), a novel phylogenetic microarray. It is designed using the PhylArray software, and is intended to assess the human gut microbiota at the family level using 4,441 25-mer probes representing 66 families present in the human gut microbiota.

Materials and Methods

Ethics statement

This study was approved by the Clinical Research Ethics Committee of the Cork Teaching Hospitals: Informed written consent was obtained from all ELDERMET subjects or, in cases of cognitive impairment, by next-of-kin in accordance with the local research ethics committee guidelines, the Clinical Research Ethics Committee of the Cork Teaching Hospitals.

Human faecal samples, bacterial strains and nucleic acids extractions

Total DNA was extracted from three human faecal samples using Qiagen's DNA Stool Kit (Qiagen, West Sussex, UK) and adjusted to 10 ng/µl. All DNA quantifications were performed using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). In order to prepare a mock community (16S rRNA bacterial amplicons), the bacterial strains *Lactobacillus acidophilus* (ATCC 4356), *Escherichia coli* (S123), *Clostridium coccoides* (ATCC 29236), *Clostridium leptum* (ATCC 29065) and *Bacteroides fragilis* (DSM 2151^{T}) were used. Total genomic DNA was extracted from pure bacterial cultures using DNeasy Blood and Tissue Kit (Qiagen, West Sussex, UK) and concentration was adjusted to 10 ng/µl to be used as 16S rRNA gene PCR amplification templates.

Microarray probe design and production

The DNA microarray was designed using a custom 16S rRNA gene database. This was derived from the sequences described in 2007 by Rajilić-Stojanović *et al.* [8] and consisted of 1,052 sequences (longer than 1,000 nucleotides) which can be accessed at http://g2im.u-clermont1.fr/HuGChip/. The PhylArray software was used to design 25-mer probes [28]. The first step of the PhylArray algorithm (Figure 1) is the extraction of all available sequences corresponding to the targeted family from our custom 16S rRNA curated database. Retrieved sequences are then aligned using the ClustalW program [30]. A degenerate consensus sequences is then deduced from this multiple alignment, taking into account the sequence variability at each position. Degenerate candidate probes are then selected along the consensus sequence, and all non-degenerate combinations are checked for cross-hybridizations against the 16S rRNA database. The locus corresponding to each 25-mer degenerate probe is referred to hereafter as a "region". Among the combinations derived from each degenerate probe, some correspond to sequences that have not yet been deposited in the databases, namely explorative probes. Such probes should, therefore, allow the detection in this environment of undescribed microorganisms belonging to the targeted taxon. The best 5 "regions" of each consensus sequence, harbouring the best specificity for the taxon were selected to represent the taxon. Finally, these selected probes were subsequently verified by BLASTN [31] against the two other databases (Greengenes [32], SILVA [33]) containing microbial sequences from many different kinds of ecosystems. The microarray was synthesized by Agilent Technologies (Agilent Technologies, Palo Alto, CA) using the *in situ* surface attached synthesis [34] with a multiplex format of 8×15k where each probe was randomly spotted in three replicates across the array to reduce biases caused by spatial variations.

16S rRNA gene PCR amplification

16S rRNA genes were amplified using universal primers 27F (AGAGTTTGATCMTGGCTCAG) and 1492R (TACGGYTACCTTGTTACGACT) [35]. PCR reactions were performed in a 50 µl volume, in the presence of 10 ng of template DNA, using DreamTaq DNA polymerase (Fermentas, St. Leon-Rot, Germany). The PCR reaction consisted of an initial denaturation step at 95°C for 5 min followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 58°C for 40 s and elongation at 72°C for 2 min. A final extension step was performed at 72°C for 5 min. PCR product size was verified by electrophoresis with 1% (w/v) agarose gel and were purified using the MinElute PCR Purification Kit (Qiagen Ltd., UK) following manufacturer's instructions and stored at −20°C. The purified amplicons from the bacterial strains were then mixed to a final amount of 1 µg of DNA composed of 100 ng of *L. acidophilus* and *E. coli*; 200 ng of *C. coccoides*; 250 ng of *B. fragilis* and 350 ng of *C. leptum* forming the mock community.

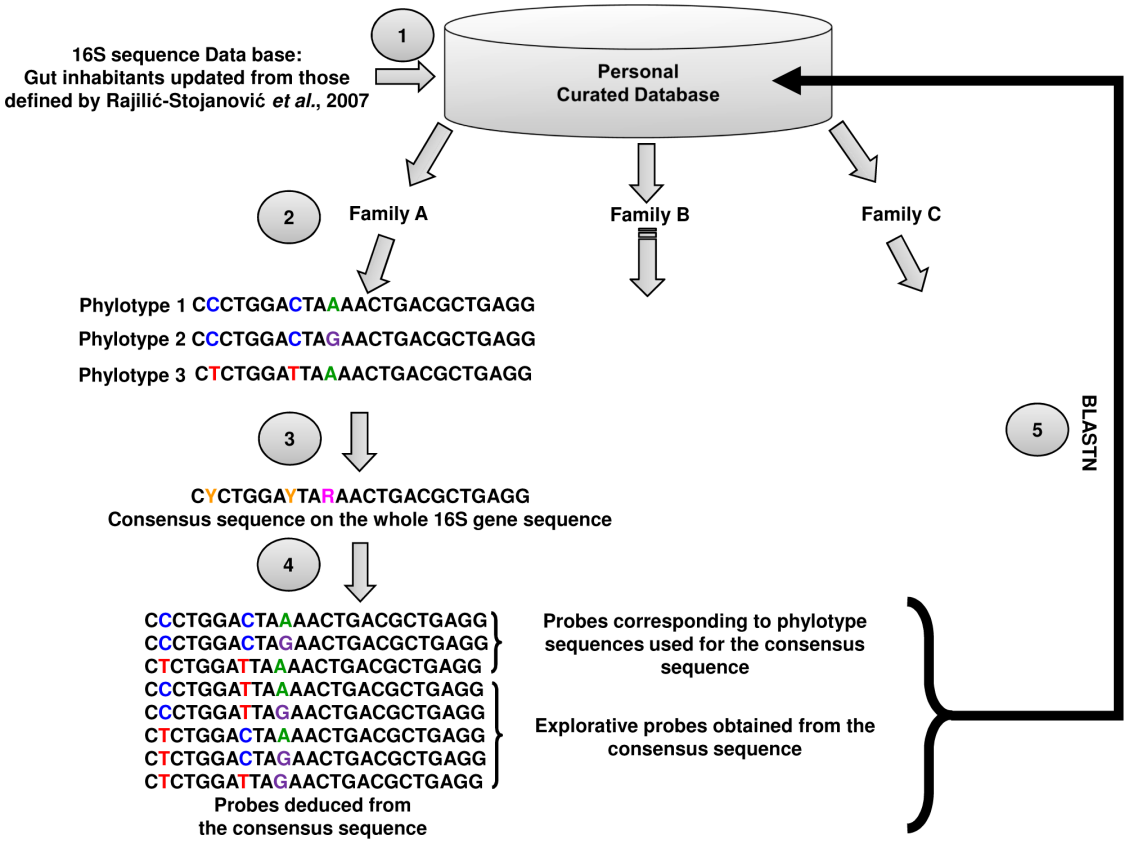


Figure 1. Probe design procedure using the PhylArray software (adapted from [27]). (1) The creation of a database was an essential part of the procedure; making sure this database contained good quality, correctly affiliated sequences was crucial. (2) The selection of a targeted taxonomic level and the reorganisation of the sequences so that they belonged to the correct taxon. (3) For each different taxon (e.g. family), a consensus sequence on the whole 16S gene sequence was constituted with all the sequences it contained. (4) The software then tested all the possible probe regions on the whole sequence using a 25 nucleotide sliding window with a step of 1 nucleotide. It selected the 5 regions with the best specificity and degeneracy for each taxon and developed all the probe combinations. (5) Finally, the software verified probe specificity performing a nucleotide BLAST against the initial database which allowed to distinct the specific from the explorative probes.
doi:10.1371/journal.pone.0062544.g001

Sample labelling and microarray hybridization, reading and analysis

For each sample (faecal samples and the mock community), the non-fragmented purified 16S rRNA gene PCR products (1 µg) were labelled with either Cy3 or Cy5 using the Genomic DNA ULS labelling Kit (Agilent Technologies, Palo Alto, CA) following the manufacturer's instructions. For microarray hybridization, 100 ng of labelled artificial bacterial DNA mix and 250 ng of each labelled faecal sample were used (GEO accession number GSE44752). Hybridization was performed following the Agilent OligoaCGH hybridization protocol (Agilent Technologies, Palo Alto, CA) at 65°C for 24 h. Microarray washings were performed as recommended by Agilent and slides were scanned at a 3-µm resolution using a Surescan microarray scanner (Agilent Technologies, Palo Alto, CA). Pixel intensities were extracted using the "Feature Extraction" software (Agilent Technologies, Palo Alto, CA). The retained intensity value for each probe was the spot's median intensity signal. For each probe, the median value of its replicates was calculated and was further identified as the "probe signal". For each of the 5 regions (considering every bacterial family), the highest probe signal was selected as the more representative probe and characterized the "region signal". For each family, a mean signal of the five "region signals" was calculated providing the "family signal". It was then used to determine the relative abundance of each family by dividing it with the sum of all the "family signals". Specific scripts developed in this study with the Delphi and the C++ languages were used to automatically perform these data extractions.

V4 16S rRNA gene pyrosequencing and metagenomic analyses of the samples

DNA extracted from three human faecal samples from the ELDERMET project (samples 176, 204 and 205) was analyzed by 454 pyrosequencing of the 16S rRNA V4 region amplicons on a 454 Genome Sequencer FLX Titanium platform as described by Claesson *et al.* [22]. Two of these samples (176 and 205) were also analysed by direct random shotgun sequencing of libraries with 91 bp paired-end Illumina reads and 350 bp insert size, further assembled using MetaVelvet [36] as described by Claesson *et al.* [5]. Raw metagenomic data are available at the MG-RAST server [37] with the following reference number 4491484.3 and 4491423.3. To determine the microbiota composition from the metagenomic samples, the rRNA sequences were affiliated using the RDP, SILVA and Greengenes database with a maximum

E-Value cut-off of $1e^{-5}$, a minimum percentage identity cut-off of 80% and a minimum alignment length cut-off of 50 nucleotides.

Quantitative PCR analysis

Quantitative PCR analysis of three phyla (*Firmicutes*, *Bacteroidetes*, *Actinobacteria*) was performed using previously published primers (Table 1) [38,39]. PCR reactions were performed in a final volume of 20 µl using Brilliant II Ultra-Fast SYBR Green qPCR Master Mix 2X (Agilent Technologies, Palo Alto, CA), in presence of 10 ng of template DNA, following the manufacturer's instructions. Quantitative PCR reactions were performed on the Mx3005P (Agilent Technologies, Palo Alto, CA). The thermocycling protocol consisted of an initial denaturation step at 95°C for 10 min followed by 40 cycles of denaturation at 95°C for 30 s, annealing at 61°C for 30 s and elongation at 72°C for 30 s, followed by a final step producing a dissociation curve. Data analysis was achieved using the Mx Pro qPCR software (Agilent Technologies, Palo Alto, CA).

Statistical analyses

Pearson correlation and one-way ANOVA with Kruskal-Wallis test and figures were performed using GraphPad Prism V 5.0 for Windows (GraphPadSoftware, San Diego, CA). Shannon's diversity index and Ward's hierarchical clustering for the samples were obtained using the Paleontological Statistics (PAST) software [40].

Results

HuGChip development and probe design

The database used for probe design was initially developed by Rajilić-Stojanović *et al.* [8] and completed to achieve a curated database of 1,052 16S rRNA gene sequences, each corresponding to a distinct phylotype. The PhylArray probe design strategy (Figure 1) was used for each family in order to take into account the sequence polymorphism (available at http://g2im.u-clermont1.fr/HuGChip/). Five non-overlapping 25-mer regions were selected within each family. For each, the number of non-degenerate combinations varied from 1 up to 182, encompassing explorative probes. Such probes should, therefore, allow the detection of undescribed microorganisms belonging to the targeted taxon. This resulted in a set of 4,441 probes (Table S1), spotted in triplicates and targeting 66 families (Table 2). The specificity of each probe was tested against the curated database: 2,442 probes were specific and 1,919 were explorative. The remaining 80 probes were redundant, meaning probes which could cross-hybridize with sequences of different families. Among them, 62 hybridized with sequences from families of the same order of the original target (Table S2). Next, the probe set was also verified using the Greengenes and SILVA databases, leading to respectively 1,852 and 1,486 specific probes. This decrease is likely due to a comparison with an exhaustive repertoire of bacterial sequences, encompassing those from families unexpected or absent in the gut environment. Among the originally defined explorative probes, only 164 and 206 had counterparts in respectively Greengenes and SILVA databases, therefore justifying the word "explorative" for all the remaining probes. The explorative probes which had counterparts in the databases were mostly specific for the intended family (respectively 141 and 136 probes accordingly to Greengenes and SILVA). The remaining 23 or 70 probes were specific for the order (Greengenes, 16 probes; SILVA, 30 probes), the class (none for Greengenes; 9 for SILVA) or the phylum (2 for Greengenes; 10 for SILVA).

In silico explorative probe assessment of the HuGChip

In order to assess the relevance of the explorative probe design strategy, these probes were tested *in silico* with metagenomic data obtained from two human faecal samples. The results indicated that 7 explorative probes could hybridize (100% identity) with metagenomic sequences, 3 with sample 176 and 4 with sample 205. As seen in Table 3, the MG-RAST affiliation of the detected sequences was in agreement with the family the probes targeted. Surprisingly, one MG-RAST affiliation was directly with a referenced strain, therefore not justifying that the probe was effectively explorative (Sequence #176-3): in fact, difference was due to the presence of ambiguous nucleotides (N) in sequences from the microarray database. Furthermore, another sequence (sequence #205-4) was detected *in silico* with a probe targeting the *Streptococcaceae* family while it was affiliated by MG-RAST as an uncultured bacterium (Table 3). When a BLASTN search was performed against the Genbank database, the best hit was with a 16S rRNA gene sequence (accession number: JX079558.1), mentioned as an uncultured *Streptococcaceae*, therefore confirming the effectiveness of this HuGChip explorative probe.

Criteria optimization for qualitative and quantitative detection of bacteria

We first decided that a bacterial family would be considered present in a sample if at least 3 of the 5 different 16S-regions showed positive signal as all the 16S rRNA regions are not accessible for hybridization in an homologous manner [41]. Then, to select the best criteria for specific detection, as well for a semi-quantitative determination of bacterial families in samples, the hybridization of a mock community of five known 16S rRNA gene amplicons was performed. This bacterial mix corresponded to 5

Table 1. Primers used for qPCR analysis of the samples.

Name	Sequence 5′-3′	Target	Annealing temp. (°C)	Source
BAC338F	ACTCCTACGGGAGGCAG	Total bacteria	61	[39]
BAC516F	GTATTACCGCGGCTGCTG			
789cfbF	CRAACAGGATTAGATACCCT	*Bacteroidetes*	61	[38]
cfb967R	GGTAAGGTTCCTCGCGTAT			
Act920F3	TACGGCCGCAAGGCTA	*Actinobacteria*	61	[38]
Act1200R	TCRTCCCCACCTTCCTCCG			
928F-Firm	TGAAACTYAAAGGAATTGACG	*Firmicutes*	61	[38]
1040FirmR	ACCATGCACCACCTGTC			

doi:10.1371/journal.pone.0062544.t001

Table 2. Phyla and families of the human gut microbiota targeted by the HuGChip.

Phylum	Family	Number of probes	Phylum	Family	Number of probes
Actinobacteria	*Actinomycetaceae*	36	Firmicutes	*Lactococcaceae*	44
	Bifidobacterium	44		*Leuconostocaceae*	36
	Coriobacteriaceae	65		*Staphylococcaceae*	10
	Corynebacteriaceae	26		*Streptococcaceae*	98
	Micrococcaceae	11		Unclassified Firmicutes	59
	Propionibacteriaceae	13		Uncultured clostridiales I-A	95
	TOTAL	***195***		Uncultured clostridiales I-B	38
Bacteroidetes	*Bacteroidaceae*	109		Uncultured clostridiales II	69
	Porphyromonodaceae A	27		**TOTAL**	***2323***
	Porphyromonodaceae B	38	Fusobacteria	*Fusobacteriaceae*	56
	Porphyromonodaceae regrouped	94		**TOTAL**	***56***
	Prevotellaceae	129	Lentisphaerae	*Victivallaceae*	5
	Rikenellaceae	49		**TOTAL**	***5***
	Uncultured Bacteroidales I	43	Proteobacteria	*Aeromonodaceae*	54
	Uncultured Bacteroidales II	19		*Alcaligenaceae*	46
	TOTAL	***508***		*Burkholderiaceae*	56
Cyanobacteria	Unclassified A	35		*Campylobacteraceae*	45
	TOTAL	***35***		*Desulfovibrionaceae*	21
Firmicutes	*Aerococcaceae*	50		*Enterobacteriaceae*	205
	Bacillaceae A	70		*Helicobacteraceae*	16
	Bacillaceae B	70		*Moraxellaceae*	35
	Bacillaceae regrouped	86		*Neisseriaceae*	117
	Carnobacteriaceae	64		*Oxalobacteriaceae*	46
	Clostridium Cluster I	115		*Pasteurellaceae*	93
	Clostridium Cluster III	28		*Pseudomonodaceae*	12
	Clostridium Cluster IV	165		*Succinivibrionaceae*	23
	Clostridium Cluster IX	198		Unclassified B	25
	Clostridium Cluster XI	127		Unclassified Rhizobiales	42
	Clostridium Cluster XIII	75		Unclassified Sphingomonadales	137
	Clostridium Cluster XIV	324		*Vibrionaceae*	102
	Clostridium Cluster XV	30		*Xanthomonodaceae*	116
	Clostridium Cluster XVI	55		**TOTAL**	***1191***
	Clostridium Cluster XVII group 1	7	Spirochaetes	*Brachyspiraceae*	12
	Clostridium Cluster XVII group 2	43		**TOTAL**	***12***
	Clostridium Cluster XVIII	86	Tenericutes	*Anaeroplasmataceae*	59
	Enterococcaceae	14		**TOTAL**	***59***
	Incertae Sedis 11	46	Verrucomicrobia	*Verrucomicrobiaceae*	57
	Lactobacillaceae	221		**TOTAL**	***57***

doi:10.1371/journal.pone.0062544.t002

different species frequently recovered from gut microbiota, in a defined ratio (Table 4). After hybridization and fluorescent signal acquisition, different signal to noise ratios (SNR) were applied to attribute a positive signal. A SNR equal or superior to 12 gave the result expected (Table 4). Furthermore, when the median of the triplicates was used and an average of the sum of the signals for each of the five regions was calculated, the relative abundance of the bacteria hybridized on the microarray was correlated to the relative abundance in the artificial bacterial mix (Pearson correlation of 0.99). Therefore, hybridization signal superior or equal to 12-fold the level of background noise indicated positive probe hybridization, *i.e.* the presence of at least one 16S-region from a bacterial family. When 3 or more regions for each family were positive with these criteria, the family was claimed present in a relative abundance defined as the mean of the signal obtained for the highest signals for each of the region used to identify the family.

Table 3. *In silico* hybridization of HuGChip explorative probes and sequences from two metagenomic samples.

Sample	Sequence ID*	HuGChip Probe	MG-RAST Affiliation
176	176-1	6947_1_10 *Bacteroidaceae*	*Bacteroides uniformis*
	176-2	6947_3_6 *Bacteroidaceae*	*Bacteroides uniformis*
	176-3	7007_1_4 *Verrucomicrobiaceae*	*Akkermansia muciniphila* ATCC BAA-835
205	205-1	6947_3_6 *Bacteroidaceae*	*Bacteroides uniformis*
	205-2	6961_3_7 Clostridium ClusterXVI	*Erysipelotrichaceae bacterium* 5_2_54FAA
	205-3	6965_4_7 *Coriobacteriaceae*	*Collinsella aerofaciens*
	205-4	6989_1_27 *Streptococcaceae*	*Uncultured bacterium*

*The Sequence IDs 176-1 to 176-3 correspond respectively to the metagenomes sequences numbers NODE_13676,NODE_30 and NODE_2236. *The Sequence IDs 205-1 to 205-4 correspond respectively to the metagenomes sequences numbers NODE_141032, NODE_71670, NODE_96151 and NODE_38960.
doi:10.1371/journal.pone.0062544.t003

Comparison of HuGChip and amplicons pyrosequencing data

DNA extracted from stool samples of 3 patients was characterized in parallel by amplicons pyrosequencing of the V4 hypervariable region of the 16S rRNA gene and the HuGChip. The results were analyzed at two different taxonomic levels, the family and the phylum level. For each taxon, the ratios of numbers of RDP classified sequence reads were compared with their corresponding relative abundance obtained with the microarray. Hierarchical clustering at family level for both techniques showed exactly the same clustering pattern (Figure S1). Following this result, Pearson's coefficients were calculated as a measurement of linear correlation between sequence-based RDP assignments ratios versus HuGChip relative abundance of all common taxonomic groups for the phylum and family (Figure 2). The results at the phylum level showed a high average Pearson's correlation coefficient (average $r = 0.92$, ranging from 0.91 to 0.94). At the family level the correlation coefficients still showed a positive correlation with an average Pearson's correlation coefficient of $r = 0.71$ (ranging from 0.63 to 0.76). The differences resulted from families which were detected by one technique but not the other: the family not detected by the HuGChip represented an average over the 3 samples of 5.6% of the total ratios, whereas the families detected by the HuGChip, but not by pyrosequencing, represented an average of 23.5% of the relative abundances. Another result was the sequences respresenting families labelled "unclassified" (e.g. unclassified Rhizobiales, unclassified Clostridiales I-A…) presented an average relative abundance varying from 18.3% to 30.2% between the HUGChip and the pyrosequencing analysis respectively. Consequently, given these results, Shannon diversity indexes were calculated showing higher indexes with the HuGChip than with pyrosequencing (Figure 3), even if considered as statistically non-significant (one way ANOVA, Kruskal-Wallis test $p = 0.062$).

Comparison of the HuGChip with metagenomic data

In order to avoid eventual bias from analyses limited to the V4 region, together with amplification bias, two of the samples mentioned above were also analyzed using random shotgun sequencing with two different levels of coverage: 14,869 sequences were obtained for the samples 176, and ~10 fold more for the sample 205 (140,766 sequences). This allowed two different sequencing depths in identified 16S rRNA features as provided by MG-RAST: 598 sequences for sample 176 and 1,458 for sample 205. The SILVA database was used to affiliate features at the phylum and family levels and results were compared to the

Table 4. Relative abundances of bacterial families at different signal to noise ratios (SNR) using a known mix of 16S rRNA amplicons.

			Relative abundances (%)				
		Amount in mix (ng)	SNR≥3	SNR≥5	SNR≥10	SNR≥12	SNR≥15
Expected Families	*Bacteroidaceae*	250	18,8	19,0	19,0	22,6	22,6
	Clostridium Cluster IV	350	28,1	28,4	28,4	33,8	33,8
	Clostridium Cluster XIV	200	17,9	18,2	18,2	21,5	21,5
	Enterobacteriaceae	100	9,7	9,9	9,9	11,7	11,7
	Lactobacillaceae	100	8,6	8,7	8,7	10,4	10,4
	Total	1000	83,1	85,2	84,2	100,0	100,0
Cross-hybridizations	*Bifidobacterium*		0,2	0,2	0,2		
	Clostridium Cluster IX		12,2	12,3	12,3		
	Coriobacteriaceae		3,2	3,3	3,3		
	Rikenellaceae		1,3				
	Total		16,9	14,8	15,8	0	0

doi:10.1371/journal.pone.0062544.t004

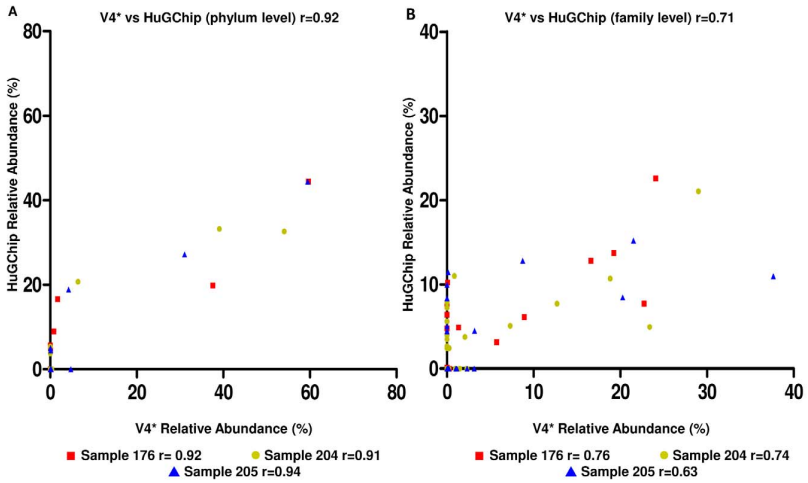


Figure 2. Comparison of relative abundances obtained with pyrosequencing (V4) and the HuGChip at two taxonomic levels. Three samples (■ 176, ● 204 and ▲ 205) were compared at both the phylum and the family level. Pearson correlation coefficients were calculated for each sample. *V4 corresponds to the pyrosequencing of the V4 hypervariable region data.
doi:10.1371/journal.pone.0062544.g002

HuGChip hybridization signals using the above criteria. Pearson correlation indicated a high similarity at both phylum and family level between the two technical approaches. As indicated in Figure 4, the average Pearson's correlation coefficient was of 0.93 at the phylum level (respectively of 0.92 and 0.94 for samples 176 and 205) and of 0.88 at the family level (respectively 0.90 and 0.85). The Greengenes and RDP databases were also used to compare the two techniques and revealed similar Pearson's correlation coefficients (data not shown). As previously, the differences relate (i) to the difficulty for the DNA microarray to detect some rare taxa, and (ii) to families detected with a relatively high abundance by the microarray which are not detected in the metagenomes. The abundance results of the three techniques were compared with qPCR for three main phyla present in the gut microbiota.

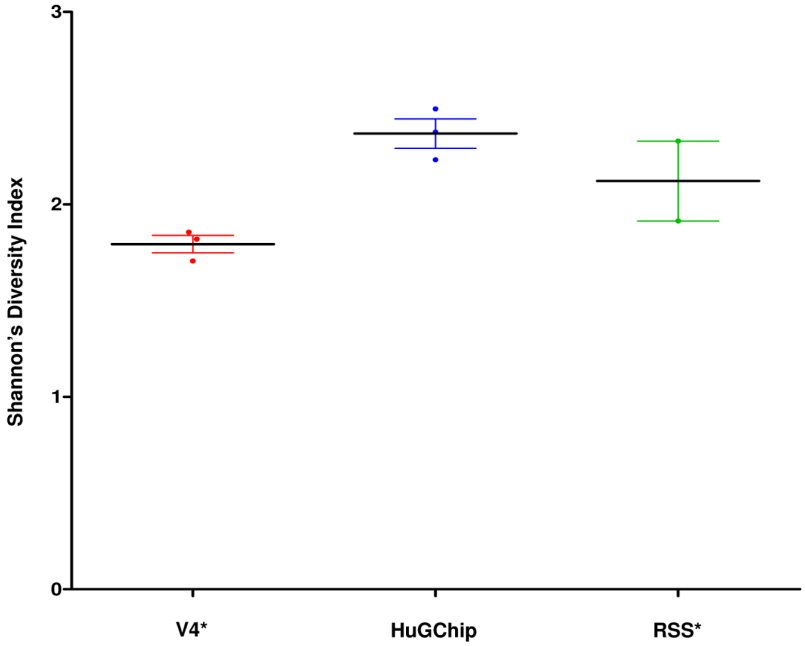


Figure 3. Comparison of Shannon's diversity index derived from the data obtained by pyrosequencing (V4), the HuGChip and metagenomics (RSS) on the faecal samples. *V4 corresponds to the pyrosequencing of the V4 hypervariable region data. **RSS corresponds to the Random Shotgun Sequencing data.
doi:10.1371/journal.pone.0062544.g003

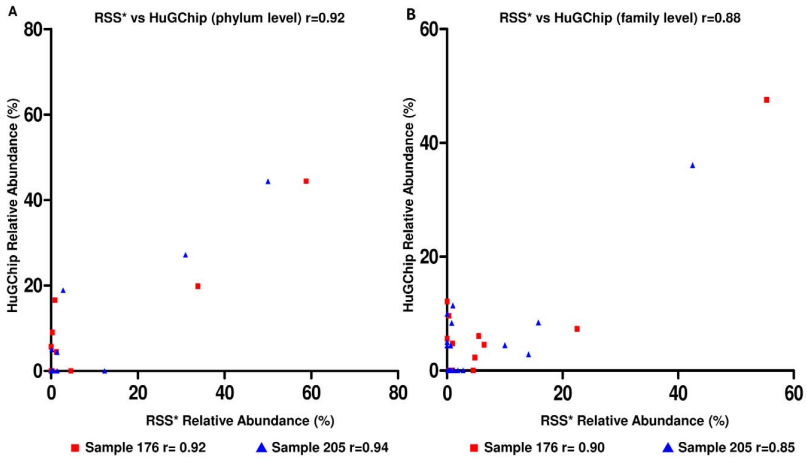


Figure 4. Comparisons of relative abundances obtained with metagenomic (RSS) and the HuGChip at two taxonomic levels. Two samples (■176 and ▲ 205) were compared at both the phylum and the family level. Pearson correlation coefficients were calculated for each sample. *RSS corresponds to the Random Shotgun Sequencing data.
doi:10.1371/journal.pone.0062544.g004

Quantitative PCR analysis and comparison with HuGChip

The qPCR technique was used here as a benchmark for quantitative analysis of the two most dominant phyla (*Firmicutes* and *Bacteroidetes*) present in faecal samples and a less abundant one (*Actinobacteria*). The results obtained confirmed that relative abundances vary slightly between the different techniques. The sequencing of the V4 region showed the highest abundances for the *Firmicutes* phylum and the HuGChip had the lowest relative abundance in only one sample (Figure 5). For the *Bacteroidetes* phylum (Figure 5), the HuGChip showed, for the three samples, the lowest abundances compared to the other techniques. Finally, it can be seen that bacterial species from the phylum *Actinobacteria* seem to be under-estimated as they are not detected with the pyrosequencing technique whereas they are detected with the three other techniques, the HuGChip giving the highest abundance (Figure 5).

Discussion

A rapid evaluation of the composition of the human gut microbiota is becoming essential in order to gain a better understanding of the interactions with the host, for example in the context of diseases, infections, ageing, or nutrition. In this study, we present a phylogenetic microarray designed at the family level that is able to assess the human gut microbiota composition. Even if differences observed between two samples at this taxonomic rank may be biologically difficult to interpret, due to functional diversity within a family, this tool should provide rapid and cheap information about the ratio of bacterial families shared among humans. This microarray was first validated *in silico*, and then optimal data interpretation regimes were empirically determined using a mock community made from reference bacterial species that inhabit the human gut. These criteria are very important as the signal to noise ratio (SNR) (as well as the number of regions considered positive) influences the qualitative and quantitative analysis of the microarray data (see Figure S2 as an example). The microarray was finally hybridized with 16S amplicons from complex samples and the results were compared with data from three other culture independent techniques applied to the same DNA samples: 454 pyrosequencing of the V4 hypervariable region of the 16S rRNA gene, metagenomic shotgun sequencing and qPCR of three selected phyla.

Microarrays are recognized as fast and user-friendly approaches to study bacterial communities [16]. Several phylogenetic microarrays have been developed to evaluate the presence and relative abundance of known bacteria from the whole human gut microbiota [8,19,42]. In contrast to other microarrays, the HuGChip, with its probe design strategy, is a phylogenetic microarray which targets known bacteria, together with potent uncharacterised respresentatives of the corresponding families. Furthermore, the design strategy allows, for each family, the determination of five regions along the 16S rRNA gene, which are not pre-defined as for example in the HITChip strategy, but are selected to give the best reliability on microarray data analysis. Twenty-five mer probes have been shown to give the best specificity [28,43] and thus, were selected for the HuGChip. Their specificities were first verified *in silico* using the sequence database used for the design indicating that the large majority of probes could be classified as specific or explorative. A small number of redundant probes were detected. These probes were frequently specific of the taxonomic levels above the family (e.g. class or order). Such hierarchical hybridization has been reported previously for other microarrays [21]. Furthermore, the probes were compared against bacterial databases containing sequences from different environments (e.g. soil, water, air, and human microbiota). The consequence was a decrease in probe specificity that might be attributed to bacterial species which had not been described in the human gut microbiota. Most of the explorative probes would not target known species, some (64 and 206 respectively for Greengenes and SILVA) could target known bacteria which were not originally detected in the human gut microbiota. Consequently, accordingly to Greengenes, only 23 of the total explorative probes were identified as hybridizing sequences from bacterial representatives from another family, including 5 targeting another phylum. These numbers rose to respectively 70 and 21 representatives using SILVA data. In fact, these results are likely over-estimates as the human gut does not host all the bacterial identified so far in all the environments. Moreover, this could explain the different relative abundance of the unclassified sequences between the pyrosequencing of the V4 hypervariable region of the 16S rRNA gene sequence and the

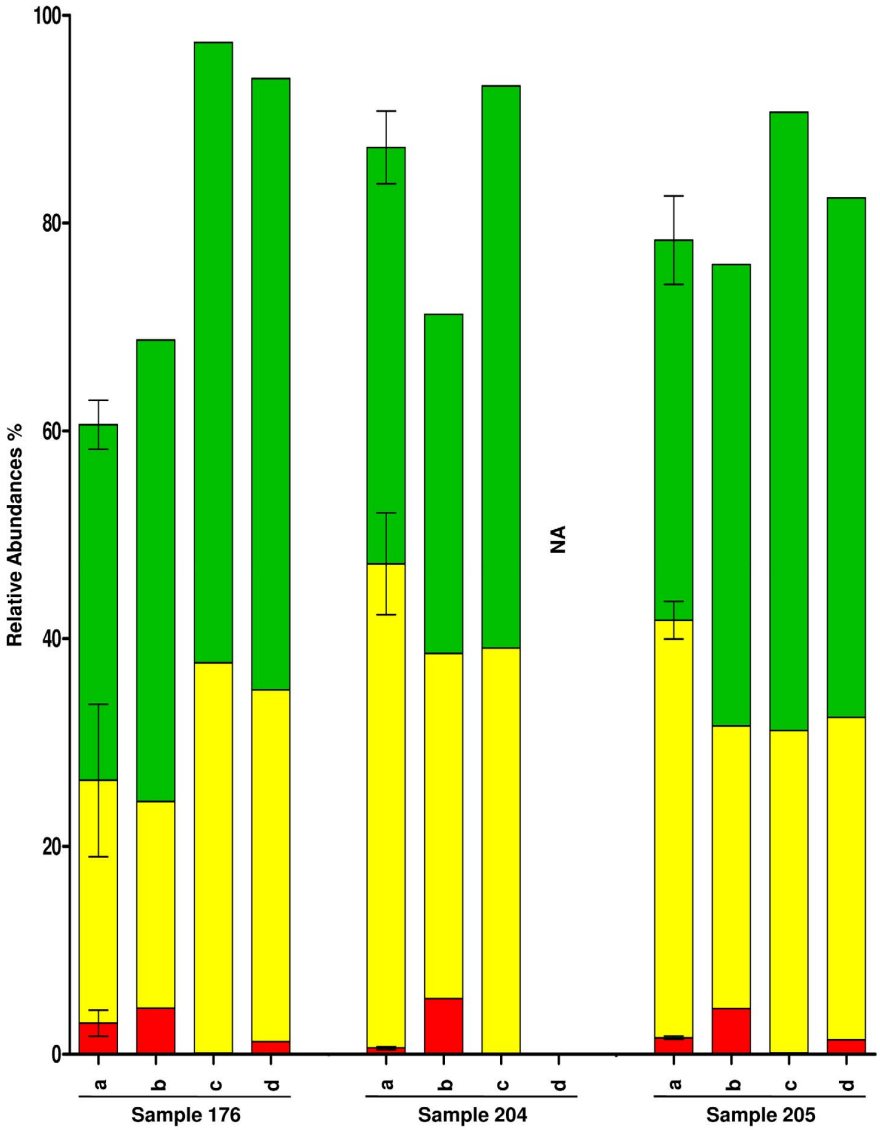


Figure 5. Comparison of qPCR results with results obtained with the HuGChip, pyrosequencing and metagenomics. The phyla *Actinobacteria* (red), *Bacteroidetes* (yellow) and *Firmicutes* (green) were analyzed by (a) qPCR (n = 3), (b) HuGChip, (c) pyrosequencing and (d) metagenomics. *NA corresponds to "not available".
doi:10.1371/journal.pone.0062544.g005

HuGChip (respectively 30,2% and 18,3%). Moreover, it was shown that 7 of the explorative probes of the HuGChip harboured 100% sequence identity and a correct taxonomic affiliation at the family level with sequences from the two metagenomes justifying their presence and benefits. These results showed that the probe design helped in minimizing the main limitations of microarrays: the detection of species which were not yet described and/or which were not included in databases used for the probe design. Other microarrays limitations could be caused by the presence of ambiguous nucleotides (N) in sequences from databases due to sequencing bias and errors: these were also at least partly overcome in this study with the use of the HuGChip explorative approach. Using this strategy, the cross-hybridization of a sequence from another family cannot be excluded but is rather unlikely and if sometimes real, contributes weakly to the overall signal, at least an order of magnitude less [20].

Using a mock community composed of 5 different families allowed setting the best threshold which had to be used with the HuGChip to analyze gut microbiota samples. As it has been shown that there are strong variations of hybridization signal intensity from probe-target duplexes with similar predicted duplexes [21,35,40,44,45], at least three of the five regions for each family have to show a probe signal to noise ratio above 12 to be considered present in the sample. These defined parameters helped to reduce the impact of possible cross-hybridizations and showed the best specificity and sensibility.

Next generation sequencing through amplicon-based or random shotgun sequencing as well as qPCR are other culture-independent techniques used to study complex ecosystems. To

further evaluate the application of the HuGChip, human faecal samples were analyzed and results were compared to these culture-independent techniques on the same samples.

Pyrosequencing of amplicons from variable regions of the 16S rRNA gene provides a deep, fast, quantitative analysis and allows the identification of unknown bacteria [4,46–49]. Although this technique specifically focuses on a hypervariable region of the 16S rRNA gene, whereas the HuGChip targets 5 regions for each family, these different approaches generated similar profiles at both the phylum and family levels. This has been already observed between the pyrosequencing of the V4 and V6 hypervariable region amplicons and the HITChip [22]. More recently, the pyrosequencing of the V1 to V6 hypervariable region amplicons of faecal and ileum lumen-content was compared with results obtained with the HITChip [23] and similar coefficients were also obtained.

Although the profiles were similar, relative abundance results between the techniques vary; it was likely due to the different means used to quantify each family, one based on sequence hit, the other on probe signal and each having their own bias [50–52]. While possible cross-hybridization or sequencing errors affect bacterial detection, incorrect or obsolete classification, annotation of sequences can also induce discrepancies. In our study, pyrosequencing of the V4 region of the 16S rRNA gene provided an important amount of unclassified sequences, part of which may have been detected and affiliated to a family due to the presence of explorative probes on the microarray. Previous studies have already shown that microarrays detected bacterial genus that were ignored by pyrosequencing of the V1 to V6 16S hypervariable regions of the 16S rRNA gene [23]. Moreover, the use of different primer sets for the HuGChip experiments and the pyrosequencing of the V4 hypervariable region may also likely contribute partly to the discrepancy observed in these two methods.

Random shotgun sequencing referred as metagenomics is another alternative culture-independent technique to study the gut microbiota, whose main advantage is the determination of large amounts of sequences from total DNA, in a more direct way, thereby avoiding PCR bias. As it does not target a particular single gene, this technique has proven to be very powerful, helping with the study of the ecosystems' metabolic potentialities and diversity [53–57]. To the best of our knowledge, this is the first time microarray data was compared to metagenomics in the perspective to address the diversity of the samples. Once again, high correlations were obtained at both phylum and family levels when the 16S gene sequences from the metagenomes were analyzed. These correlations were equivalent or even higher than the coefficients obtained between pyrosequencing and the microarray. The minor differences observed between the two techniques were certainly attributed to 2 congruent reasons: the microarray's sample preparation procedure (necessitating PCR, and consequently a potent quantitative bias) and the low number of ribosomal sequences available for taxonomic attributions from the metagenomic results (around 1,500 for the deepest sequenced sample).

The results of the three techniques were finally compared to qPCR at the phylum level. This is a commonly used technique to quantify specific taxonomic groups in a sample. Even if differences were seen among the techniques for the three phyla tested, they were likely due to the low number of experiments and that all the techniques present globally similar abundance patterns. The microarray gave a higher signal for the low-represented phylum (*Actinobacteria*) compared to 16S pyrosequencing and metagenomics, near to qPCR values. Taken into account that primers used in this study to amplify 16S rRNA gene sequences of the samples should rather lead to an underestimation of *Bifidobacterium spp* from the phylum *Actinobacteria*, it remains to be determined whether this is due to this particular taxonomic group or to the fact that it corresponds to a low-represented phylum, which is under-detected with pyrosequencing methods. Taken into account that the HuGChip gave higher Shannon Diversity Index when compared with either 16S pyrosequencing or metagenomics argues preferentially for a better evaluation of low-represented families while dominant ones (*Firmicutes* and *Bacteroidetes*) seemed to be less prevailing.

Altogether, the results showed that the HuGChip is a suitable tool to assess the human gut microbiota. Contrary to other microarrays, this tool contains explorative probes which allow the detection of unknown bacteria, without providing strong taxonomic evidences, but probably contributes to a better detection of low-represented families, and increases the specificity at the family level thanks to the use of 5 different regions per family. Pyrosequencing of the V4 region of the 16S rRNA gene provided an important amount of unclassified sequences, part of which may have been detected and affiliated by the microarray to a family: in fact, the presence of explorative probes based on 5 specific "regions" spread along the 16S rRNA gene and not restricted to a small variable region is a significant improvement as a majority of the explorative probes do not show counterparts in international database used for the affiliation of sequencing data. This suggests also that the microarray could be used for other environments, in which bacterial families are similar: this encompasses samples from other compartments of the digestive tract that have different bacterial compositions [49] and that partially explain the discrepancies between the HITChip and pyrosequencing of the V1 to V6 16S hypervariable regions observed in a previous study [23]. This might be avoided by using the HuGChip, which could evaluate the microbiota from these different compartments in the human host, but also in other animals (e.g. rodents, ruminants).

In this study, we showed that the HuGChip had similar profiles at both the phylum and the family level. This microarray can thus be considered as a suitable tool to analyze the human gut microbiota as it is a rapid, cheap and user friendly technique which allows studying several samples in parallel. Currently, the format and design of the HuGChip (8×15k probes, three probe replicates) make it possible to analyze 16 different samples per run reducing costs and limiting inter microarray bias. Furthermore, the analysis of the data extracted from the microarray is not laborious compared to other high throughput techniques and stands on 5 different regions per family, increasing specificity. Microarrays are also a particularly well-adapted format to monitor the gut bacterial environment over the time and are a mean to give an alternative determination of the bacterial richness and abundance of a sample. Taken altogether, this suggests that the microarray should also be used to characterize and select the samples of interests in order to study them with next generation sequencing techniques. Especially, improved techniques such as MiSeq Illumina technology or emerging third generation sequencing which may bring increased depth of analysis with lower time of analysis, and will surely provide new knowledge of the gut microbiota's composition, structure and role within the human health.

Supporting Information

Figure S1 Impact of threshold selection on the results of a complex sample.
(PPTX)

Figure S2 Euclidean clustering of the three samples when they are analyzed by (a) pyrosequencing and (b) the HuGChip.
(PPTX)

Table S1 Targeted family, sequence, region and localization on the 16S rRNA gene of the probes spotted on the HuGChip.
(XLSX)

Table S2 Greengene Database (July 2011).
(XLSX)

Acknowledgments

We thank the team of Philippe Langela (INRA of Jouy-en-Josas), Adeline Regnier and Emilie Girard for their technical help.

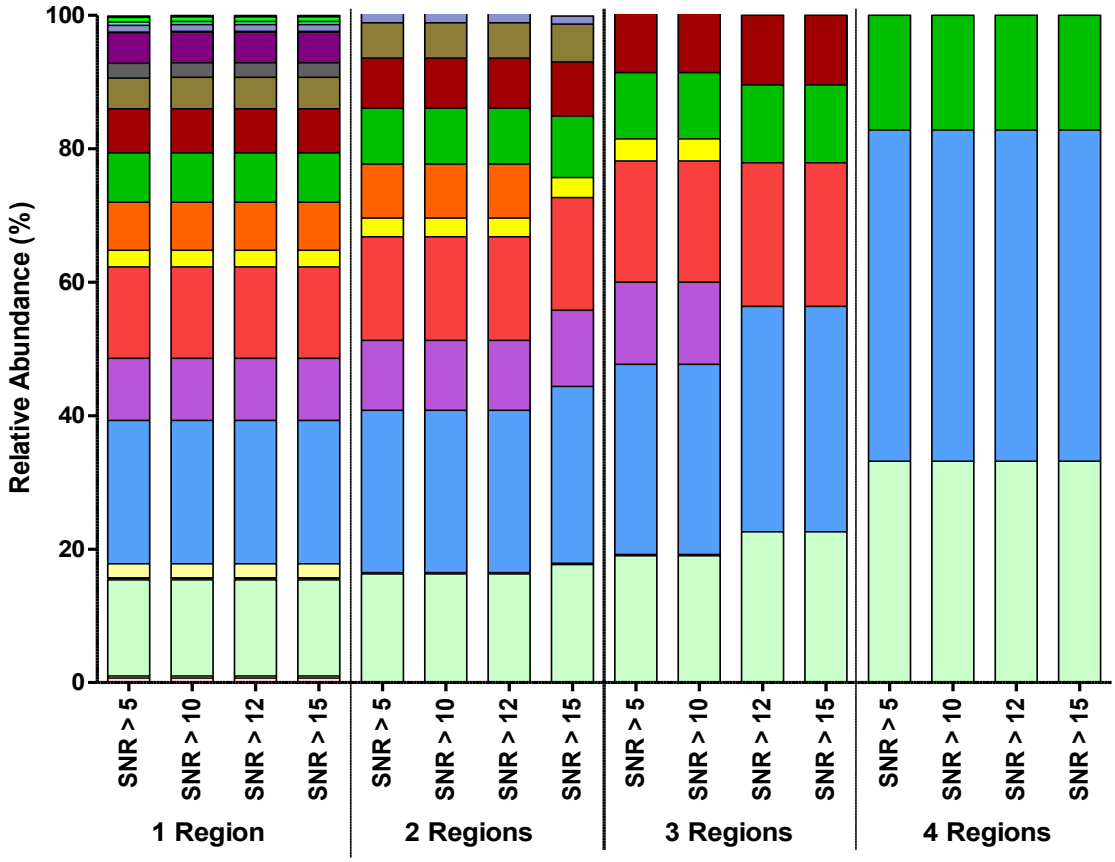
Author Contributions

Conceived and designed the experiments: WT JD MM DH EP MA PWOT PP JFB. Performed the experiments: WT JD GB IBJ MJC JFB. Analyzed the data: WT FJ NP HMBH IBJ MJC. Wrote the paper: WT JD PWOT PP JFB.

References

1. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annual Reviews in Microbiology 31: 107–133.
2. Mihajlovski A, Alric M, Brugère JF (2008) A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the *mcrA* gene. Research in microbiology 159: 516–521.
3. Scanlan PD, Shanahan F, Marchesi JR (2008) Human methanogen diversity and incidence in healthy and diseased colonic groups using mcrA gene analysis. BMC microbiology 8: 79.
4. Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, De Weerd H, et al. (2011) Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proceedings of the National Academy of Sciences 108: 4586.
5. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, et al. (2012) Gut microbiota composition correlates with diet and health in the elderly. Nature 488: 178–184.
6. Mihajlovski A, Doré J, Levenez F, Alric M, Brugère JF (2010) Molecular evaluation of the human gut methanogenic archaeal microbiota reveals an age-associated increase of the diversity. Environmental Microbiology Reports 2: 272–280.
7. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, et al. (2009) Towards the human intestinal microbiota phylogenetic core. Environmental microbiology 11: 2574–2584.
8. Rajilić-Stojanović M, Smidt H, De Vos WM (2007) Diversity of the human gastrointestinal tract microbiota revisited. Environmental microbiology 9: 2125–2136.
9. Clemente JC, Ursell LK, Parfrey LW, Knight R (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148: 1258–1270.
10. Hill M (1997) Intestinal flora and endogenous vitamin synthesis. European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP) 6: S43.
11. McGarr SE, Ridlon JM, Hylemon PB (2005) Diet, anaerobic bacterial metabolism, and colon cancer: a review of the literature. Journal of clinical gastroenterology 39: 98.
12. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.
13. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 55: 205–211.
14. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences 104: 13780.
15. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, et al. (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472: 57–63.
16. Fraher MH, O'Toole PW, Quigley EMM (2012) Techniques used to characterize the gut microbiota: a guide for the clinician. Nature Reviews Gastroenterology and Hepatology 9: 312–322.
17. Brugère JF, Mihajlovski A, Missaoui M, Peyret P (2009) Tools for stools: the challenge of assessing human intestinal microbiota using molecular diagnostics. Expert Review of Molecular Diagnostics 9: 353–365.
18. DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, et al. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microbial Ecology 53: 371–383.
19. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO (2007) Development of the human infant intestinal microbiota. PLoS biology 5: e177.
20. Paliy O, Kenche H, Abernathy F, Michail S (2009) High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray. Applied and environmental microbiology 75: 3572–3579.
21. Rajilić-Stojanović M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, et al. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environmental microbiology 11: 1736–1751.
22. Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, et al. (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 4: e6669.
23. Van Den Bogert B, De Vos WM, Zoetendal EG, Kleerebezem M (2011) Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. Applied and environmental microbiology 77: 2071–2080.
24. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Research 32: 1363–1371.
25. Ashelford KE, Weightman AJ, Fry JC (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. Nucleic Acids Research 30: 3481–3489.
26. Severgnini M, Cremonesi P, Consolandi C, Caredda G, De Bellis G, et al. (2009) ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. Nucleic Acids Research 37: e109–e109.
27. Dugat-Bony E, Peyretaillade E, Parisot N, Biderre-Petit C, Jaziri F, et al. (2012) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. Environmental microbiology 14(2): 356–371.
28. Militon C, Rimour S, Missaoui M, Biderre C, Barra V, et al. (2007) PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics 23: 2550–2557.
29. Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E (2012) KASpOD-A web service for highly specific and explorative oligonucleotide design. Bioinformatics 28(23): 3161–3162.
30. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403–410.
32. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72: 5069–5072.
33. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35: 7188–7196.
34. Blanchard A, Kaiser R, Hood L (1996) High-density oligonucleotide arrays. Biosensors and Bioelectronics 11: 687–690.
35. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. Journal of bacteriology 173: 697–703.
36. Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40(20): e155.
37. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.
38. Bacchetti De Gregoris T, Aldred N, Clare AS, Burgess JG (2011) Improvement of phylum-and class-specific primers for real-time PCR quantification of bacterial taxa. Journal of microbiological methods 86: 351–356.
39. Yu Y, Lee C, Kim J, Hwang S (2005) Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. Biotechnology and bioengineering 89: 670–679.
40. Hammer Ø, Haprper DAT, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. Available: http://www.uv.es/pe/2001_1/past/past.pdf. Accessed 2013 Apr 25.
41. Peplies J, Glöckner FO, Amann R (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Applied and environmental microbiology 69: 1397–1407.
42. Paliy O, Agans R (2012) Application of phylogenetic microarrays to interrogation of human microbiota. FEMS microbiology ecology 79: 2–11.
43. Harrington CR, Lucchini S, Ridgway KP, Wegmann U, Eaton TJ, et al. (2008) A short-oligonucleotide microarray that allows improved detection of gastrointestinal tract microbial communities. BMC microbiology 8: 195.
44. Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, et al. (2006) Rapid quantitative profiling of complex microbial populations. Nucleic Acids Research 34: e5–e5.

45. Bodrossy L, Stralis-Pavese N, Murrell JC, Radajewski S, Weilharter A, et al. (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. Environmental microbiology 5: 566–582.
46. Davis LMG, Martínez I, Walter J, Goin C, Hutkins RW (2011) Barcoded Pyrosequencing Reveals That Consumption of Galactooligosaccharides Results in a Highly Specific Bifidogenic Response in Humans. PLoS One 6: e25200.
47. LaTuga MS, Ellis JC, Cotton CM, Goldberg RN, Wynn JL, et al. (2011) Beyond Bacteria: A Study of the Enteric Microbial Consortium in Extremely Low Birth Weight Infants. PLoS One 6: e27858.
48. Roh SW, Abell GCJ, Kim KH, Nam YD, Bae JW (2010) Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends in biotechnology 28: 291–299.
49. Stearns JC, Lynch MDJ, Senadheera DB, Tenenbaum HC, Goldberg MB, et al. (2011) Bacterial biogeography of the human digestive tract. Scientific Reports 1. doi:10.1038/srep00170.
50. Smith CJ, Osborn AM (2008) Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. FEMS microbiology ecology 67: 6–20.
51. Gentry T, Wickham G, Schadt C, He Z, Zhou J (2006) Microarray applications in microbial ecology research. Microbial Ecology 52: 159–175.
52. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clinical chemistry 55: 856–866.
53. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
54. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. science 312: 1355–1359.
55. Lepage P, Leclerc MC, Joossens M, Mondot S, Blottière HM, et al. (2012) A metagenomic insight into our gut's microbiome. Gut 62(1): 146–158.
56. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2008) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
57. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.



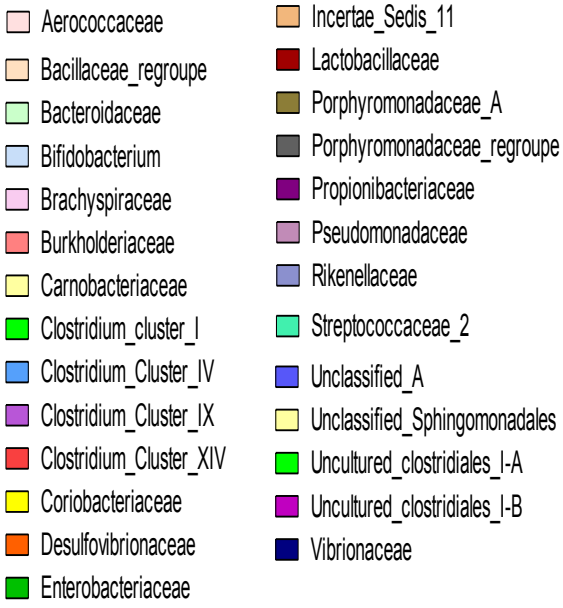


Figure S1: impact of threshold Selection on the results of a Complex sample.

Signal to noise ratio and the number of regions were tested showing important variations.

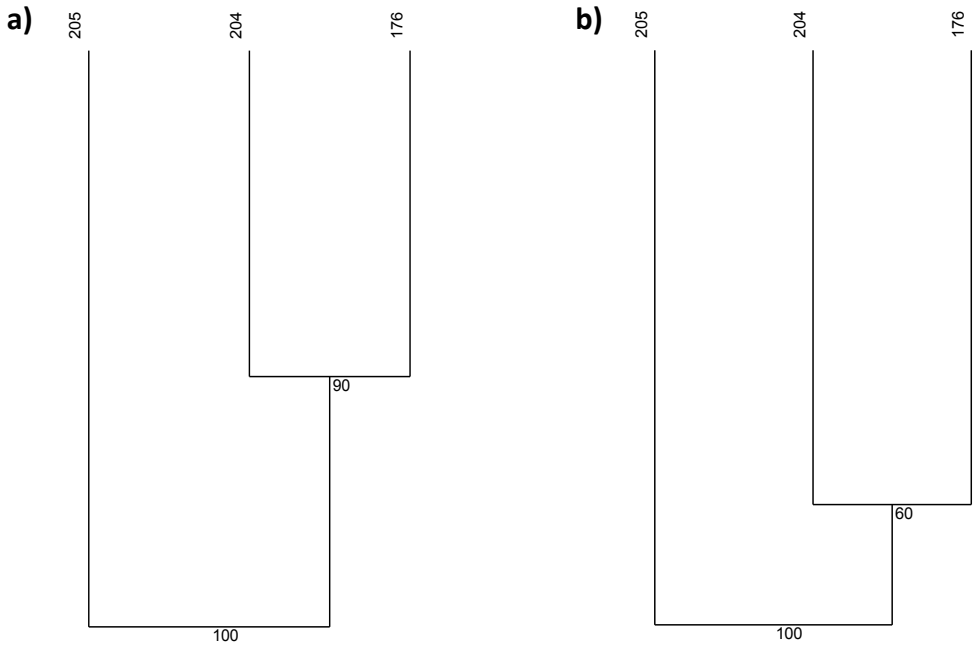


Figure S2: Euclidean Clustering of the three samples when they are analyzed by a) pyrosequencing and b) the HuGChip

1.3.2 Biopuce phylogénétique généraliste

Grâce aux développements bioinformatiques conjoints des logiciels KASpOD et PhylGrid 2.0, deux jeux de sondes oligonucléotidiques ciblant le gène ADNr 16S ont pu être obtenus. KASpOD a permis la détermination de 56 613 sondes ciblant 1295 genres procaryotes alors que PhylGrid 2.0 a assuré la sélection de 19 874 oligonucléotides supplémentaires permettant l'étude de 2069 genres procaryotes. L'ensemble des sondes ainsi obtenues a été utilisé pour la production d'une biopuce phylogénétique sur une plateforme *Agilent Technologies* 2×400k (*i.e.* 5 réplicats pour chaque sonde et 4 échantillons analysés en parallèle).

Cette biopuce est en cours de validation biologique et les premiers résultats tendent à montrer qu'il existe une complémentarité des deux jeux de sondes plutôt qu'une supériorité de l'un par rapport à l'autre.

1.4 Discussion

Ces travaux visent au développement de nouvelles biopuces phylogénétiques adaptées aux problématiques d'écologie microbienne. A travers leur caractère exploratoire, elles revêtent un intérêt indéniable pour l'étude des microorganismes, caractérisés ou non, au sein de divers environnements. La biopuce HuGChip constitue à l'heure actuelle la seule biopuce ADN exploratoire capable d'explorer les communautés microbiennes du tractus intestinal humain en proposant une précision d'analyse et un débit concurrentiels des approches par séquençage.

En effet, le séquençage massif et les biopuces ADN sont actuellement les deux techniques haut-débit les plus prometteuses et complémentaires pour l'étude des communautés microbiennes au sein d'environnements complexes. De nombreuses études, y compris celle menée pour la validation de la HuGChip, ont permis de montrer que les résultats de ces deux stratégies étaient fortement corrélés (Claesson *et al.* 2009 ; Roh *et al.* 2010 ; van den Bogert *et al.* 2011 ; Tottey *et al.* 2013). Même si les biopuces ADN ne permettent pas d'avoir accès aux séquences, il est possible de citer plusieurs avantages pour l'identification bactérienne, comme la possibilité d'utiliser plusieurs sondes réparties le long du gène ADNr 16S afin d'améliorer la précision d'affiliation contrairement au séquençage d'amplicons qui ne se focalise que sur une portion restreinte du gène. Par ailleurs, les biopuces ADN permettent de s'affranchir des limites engendrées par la PCR en permettant

l'hybridation directe des échantillons extraits puis marqués. Une approche métagénomique couplée au séquençage massif permettrait d'obtenir des résultats similaires mais elle nécessiterait un effort de séquençage trop important pour la rendre applicable à l'étude d'un environnement aussi complexe que le microbiote colique humain. En effet, les biopuces ADN sont capables de détecter des populations très peu abondantes (*i.e.* entre 0,03% et 5% de la communauté totale (Palmer *et al.* 2006 ; Rajilić-Stojanović *et al.* 2009)) qui peuvent être non identifiées par les approches de séquençage partielles (Quince *et al.* 2008).

Néanmoins, les stratégies de réduction de complexité, telles que la capture de gènes couplée au séquençage haut-débit, permettent de s'affranchir des limites évoquées précédemment en enrichissant de manière significative l'échantillon étudié en séquences d'intérêt.

2. Développement d'une méthode innovante de capture de gènes en solution couplée à du séquençage haut-débit pour l'exploration métagénomique ciblée des environnements complexes

2.1 Contexte

L'émergence des nouvelles techniques de séquençage (NGS) permet à l'heure actuelle d'étudier directement l'ADN total extrait d'un environnement (métagénome) sans passer par la construction de banques de clones nécessaire au séquençage par la méthode de Sanger (Edwards *et al.* 2006). Ces NGS (Shendure & Ji 2008 ; Ansorge 2009 ; Metzker 2010 ; Glenn 2011) offrent de nouvelles opportunités pour explorer et étudier les communautés microbiennes jusqu'alors non cultivées et non caractérisées au sein des environnements complexes (Venter *et al.* 2004 ; Sogin *et al.* 2006 ; Eisen 2007 ; Claesson *et al.* 2010 ; Caporaso *et al.* 2011 ; Shokralla *et al.* 2012).

Cependant, une exploration des environnements complexes dans leur globalité, nécessite un effort de séquençage très important, dépassant les capacités actuelles des NGS (Quince *et al.* 2008). De plus, la quantité importante de données générées, la longueur des lectures encore limitée (de 20 bases à 1 kb avec le développement récent des NGS de troisième génération) ou le taux d'erreurs de séquençage restent des problèmes majeurs notamment pour assurer l'assemblage des séquences et permettre la reconstruction de génomes ou de grandes régions d'ADN (Hoff 2009). A l'heure actuelle, l'utilisation de ces nouvelles technologies reste encore limitée pour explorer finement les environnements complexes et coûteuse pour de nombreuses structures de recherche (Bentley 2006 ; Roh *et al.* 2010). Une alternative intéressante serait donc de pouvoir réduire la complexité des échantillons métagénomiques sans utiliser la PCR, source importante de biais, en enrichissant spécifiquement les séquences nucléiques d'intérêt.

2.2 Objectif

Afin de proposer une nouvelle alternative en écologie microbienne pour l'étude des environnements complexes en lien avec l'essor des nouvelles méthodes de séquençage, l'objectif de ce travail a été de développer une nouvelle méthode de capture de gènes, utilisant des sondes sensibles, spécifiques et exploratoires, combinée au séquençage de deuxième

génération. Actuellement, aucune approche de capture de gènes utilisant des sondes n'a été appliquée sur des échantillons métagénomiques. Cette méthode, basée sur la capture de gènes en solution (Gnirke *et al.* 2009), représente une nouvelle approche moléculaire en écologie microbienne permettant de réduire la complexité des métagénomes étudiés et donc d'assurer une exploration ciblée des communautés microbiennes d'intérêt. Cette approche, tout en permettant d'explorer de manière exhaustive la diversité de gènes d'intérêt, présente l'avantage d'assurer l'obtention d'une plus grande portion, voire l'intégralité, du gène d'intérêt garantissant une résolution d'analyse importante. Il est aussi possible, par la caractérisation de grandes régions d'ADN, d'identifier les régions flanquantes associées aux séquences ciblées pour mettre en évidence de nouvelles organisations génomiques voire reconstruire de nouveaux opérons et donc identifier de nouveaux gènes pouvant avoir un rôle dans une voie métabolique donnée. Il faut également noter que contrairement à la PCR qui nécessite l'identification de deux régions conservées pour définir deux séquences oligonucléotidiques, une seule peut être suffisante pour cette approche.

Afin d'évaluer l'efficacité de cette nouvelle méthode, elle a tout d'abord été appliquée en ciblant et en enrichissant le gène codant pour la méthyl-coenzyme M réductase (*mcrA*) directement à partir du génome de la souche *Methanosarcina acetivorans* C2A. Puis elle a été utilisée pour explorer la diversité des communautés méthanogènes au niveau de la zone anoxique d'un lac méromictique. Par la suite, elle a été transposée à l'étude d'autres environnements et d'autres biomarqueurs.

2.3 Principaux résultats

Les travaux ont conduit à la rédaction d'une publication dans le journal « *DNA Research* ». Cette étude s'est inscrite dans une problématique méthodologique, c'est-à-dire proposer un outil efficace et pertinent pour l'étude de la diversité des microorganismes des environnements complexes, et également biologique en relation avec la production de méthane.

La validation de la méthode a été réalisée premièrement en enrichissant spécifiquement le gène *mcrA* (~1,6 kpb) au sein du génome de la souche *Methanosarcina acetivorans* C2A (~5,8 Mpb) en utilisant un jeu de six sondes déterminées par le logiciel HiSpOD (Dugat-Bony *et al.* 2011) et ciblant différentes régions du gène. Suite à la capture, les cibles ont été clonées puis séquencées et analysées par qPCR. Après deux cycles de capture 100% des séquences piégées et séquencées correspondent au gène *mcrA*. Ces résultats sont confirmés par

l'approche de PCR quantitative qui montre un enrichissement d'un facteur 461 et 175 365 respectivement pour le premier et le deuxième cycle de capture. Ces résultats traduisent l'efficacité de la méthode pour enrichir spécifiquement le gène *mcrA* à partir du génome de la souche étudiée. Une deuxième validation a été réalisée en utilisant un jeu de 26 sondes, définies par le logiciel HiSpOD (Dugat-Bony *et al.* 2011), ciblant toute la diversité du gène *mcrA*/*mrtA* présente dans les bases de données et permettant l'identification de variants géniques encore non identifiés. Ces sondes ont été utilisées pour étudier l'ADN métagénomique extrait de la zone anoxique du lac Pavin abritant des communautés d'archées méthanogènes. Sur les dix fragments capturés et clonés, cinq correspondent au gène *mcrA* avec des similarités significatives (99%) avec des séquences nucléotidiques isolées auparavant par PCR au sein de ce même écosystème. De plus, ces séquences ont permis d'avoir accès aux régions flanquantes du gène *mcrA,* avec des portions couvrant les gènes *mcrG* et *mcrC* (gènes de l'opéron codant pour la méthyl-coenzyme M réductase) ou mettant en évidence un gène (*fmdC*) adjacent à l'opéron *mcr* et impliqué dans la méthanogénèse hydrogénotrophe. Ces résultats mettent en avant le potentiel de l'approche pour enrichir significativement l'ADN métagénomique en séquence *mcrA*, mais également pour capturer de grandes régions génomiques permettant d'explorer les régions adjacentes du gène ciblé.

Après avoir montré la pertinence de l'approche capture de gènes, celle-ci a été comparée à une approche métagénomique directe et à une approche PCR utilisant des amorces universelles du gène *mcrA*. Environ 100 000 lectures pour chaque approche ont été générées puis traitées pour être au final regroupées au sein d'OTUs à un seuil de 91% au niveau protéique. Une diversité totale de 58 OTUs a été observée avec seulement 1 OTU provenant de l'approche métagénomique directe, 40 OTUs de l'approche PCR et 44 OTUs de l'approche capture. L'analyse phylogénétique a montré que toutes les séquences identifiées par l'approche PCR étaient affiliées au niveau de trois ordres différents alors que l'approche capture a permis de caractériser des séquences correspondant à ces trois mêmes ordres, mais également à celui des *Methanobacteriales*. Ces résultats montrent donc une évaluation plus exhaustive de la diversité par l'approche capture en comparaison avec l'approche PCR. De plus, une approche d'assemblage des séquences issues de la capture a permis de reconstruire des *contigs* permettant d'explorer les régions flanquantes du gène *mcrA*. Il a ainsi été possible d'identifier la séquence de gènes adjacents mais également de mettre en évidence une organisation génétique encore jamais décrite chez les archées méthanogènes. Ces résultats soulignent la pertinence de l'approche de capture de gènes pour explorer de manière ciblée la

diversité des communautés microbiennes, et ceci de manière plus complète que ne le permettent les approches moléculaires classiques comme la PCR. Grâce à cette étude, il a pu aussi être montré que l'approche peut être facilement couplée à des approches de séquençage massif pour évaluer la diversité totale d'un écosystème et/ou pour assurer la reconstruction de grandes régions génomiques. Celles-ci peuvent mettre en lumière de nouvelles organisations géniques traduisant éventuellement l'existence d'adaptations métaboliques particulières chez les microorganismes étudiés ou d'identifier de nouveaux gènes potentiellement impliqués dans les voies métaboliques ciblées.

Article n°7

Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration.

DNA Research **20**, 185–196, (2013) doi:10.1093/dnares/dst001
Advance Access publication on 30 January 2013

Gene Capture Coupled to High-Throughput Sequencing as a Strategy for Targeted Metagenome Exploration

Jérémie Denonfoux[1,2,3,†], Nicolas Parisot[1,2,3,†], Eric Dugat-Bony[1,4], Corinne Biderre-Petit[3,5], Delphine Boucher[1,4], Diego P. Morgavi[6], Denis Le Paslier[7,8,9], Eric Peyretaillade[1,4], and Pierre Peyret[1,4,*]

Centre de Recherche en Nutrition Humaine Auvergne, Clermont Université, Université d'Auvergne, EA 4678, Conception, Ingénierie et Développement de l'Aliment et du Médicament, BP 10448, Clermont-Ferrand 63000, France[1]; Clermont Université, Université Blaise Pascal, Clermont-Ferrand 63000, France[2]; UMR CNRS 6023, Université Blaise Pascal, Clermont-Ferrand 63000, France[3]; UFR Pharmacie, Clermont Université, Université d'Auvergne, Clermont-Ferrand 63000, France[4]; Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, Clermont-Ferrand 63000, France[5]; INRA, UMR1213 Herbivores, F-63122 Saint-Genès-Champanelle and Clermont Université, VetAgro Sup, UMR Herbivores, BP 10448, F-63000, Clermont-Ferrand, France[6]; CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, Evry 91057, France[7]; CNRS, UMR8030, Evry 91057, France[8] and UEVE, Université d'Evry, Evry 91057, France[9]

*To whom correspondence should be addressed. EA4678 CIDAM, 28 place Henri Dunant, 63001 Clermont-Ferrand. Tel. +33 47-3178-308. Fax. +33 47-3275-624. Email: pierre.peyret@udamail.fr

Edited by Prof. Masahira Hattori
(Received 10 October 2012; accepted 9 January 2013)

Abstract

Next-generation sequencing (NGS) allows faster acquisition of metagenomic data, but complete exploration of complex ecosystems is hindered by the extraordinary diversity of microorganisms. To reduce the environmental complexity, we created an innovative solution hybrid selection (SHS) method that is combined with NGS to characterize large DNA fragments harbouring biomarkers of interest. The quality of enrichment was evaluated after fragments containing the methyl coenzyme M reductase subunit A gene (*mcrA*), the biomarker of methanogenesis, were captured from a *Methanosarcina* strain and a metagenomic sample from a meromictic lake. The methanogen diversity was compared with direct metagenome and *mcrA*-based amplicon pyrosequencing strategies. The SHS approach resulted in the capture of DNA fragments up to 2.5 kb with an enrichment efficiency between 41 and 100%, depending on the sample complexity. Compared with direct metagenome and amplicons sequencing, SHS detected broader *mcrA* diversity, and it allowed efficient sampling of the rare biosphere and unknown sequences. In contrast to amplicon-based strategies, SHS is less biased and GC independent, and it recovered complete biomarker sequences in addition to conserved regions. Because this method can also isolate the regions flanking the target sequences, it could facilitate operon reconstructions.

Key words: α-subunit of the methyl-coenzyme M reductase; metagenomics; sequence capture; 454 pyrosequencing; microbial diversity

1. Introduction

Microorganisms are extremely diverse and crucial for healthy, functioning biospheres.[1,2] Although studies of isolated species have produced a great deal of information about microbial genetics, physiology, biotechnology and molecular biology, the diversity and structure of complex microbial communities are still poorly understood. This deficiency results from the inability to culture most microorganisms using standard microbiological techniques.[1,3] Consequently, although there are most likely millions of bacterial

† These authors contributed equally to this study.

species on the planet, only a few thousand have been formally described.[4]

Culture-independent techniques, such as metagenomics,[5] circumvent the problem of unculturability and transcend previous studies on individual organisms to focus on microbial communities present in an environment. Metagenomics has enriched our knowledge of environmental microbiology through the structural (gene/species richness and distribution)[6] and functional (metabolic)[7] profiling of complex environmental microbial communities. Based on unselective (shotgun) or targeted (activity driven and sequence driven) methods, metagenomics links genome information with structure and function relationships within microbial populations.[8,9]

Recently developed next-generation sequencing (NGS) technologies recover genetic materials from environmental samples without the preparation of metagenomic clone libraries.[10] Furthermore, they explore a greater amount of sequence information because they have higher throughput and lower costs than other methods.[11] Nevertheless, Quince *et al.*[12] showed that covering 90% of the species richness in some hyper-diverse environments could require 10—1000-fold increases in the current NGS sequencing efforts. In addition, the massive amount of short metagenomic sequence reads (between 20 and 700 bases depending on the platform) can be problematic for assembling and identifying complete coding DNA sequence and/or operon structure.[13] One promising alternative is to reduce the environmental sample complexity by enriching the desired genomic target before sequencing.

Currently, several strategies of genomic-scale sequence enrichment have been reported.[14] The more efficient methods rely on complementary nucleic acid capture probes that hybridize to the targeted DNA sequences. Two hybridization methods—solid phase[15–17] and solution phase, also known as solution hybrid selection (SHS)[18,19]—can be used to ascertain genetic variation by specifically enriching and resequencing regions from complex eukaryotic genomes.

To the best of our knowledge, only high-throughput enrichment methods based on polymerase chain reaction (PCR) have been applied to target functional genes in complex environments.[20] Because no current methods use oligonucleotide capture probes to specifically enrich targeted genes from a complex environmental genomic DNA (gDNA), we applied this methodology in the context of microbial ecology (Fig. 1A) to specifically capture DNA fragments harbouring known or unknown genetic biomarkers of interest (Fig. 1B). We hypothesized that the use of variant specific and explorative probes[21,22] would more accurately define the overall biomarker diversity (including the rare biosphere and unknown sequences) and would facilitate the discovery of genes linked to the target sequences *via* the reconstruction of adjacent DNA regions. This method should lead to better diversity coverage that is not influenced by PCR biases, as generally occurs in amplicon sequencing.[23,24] Because it is not limited to a specific DNA region (as in PCR enrichment), this strategy will increase the sequence coverage over target regions and lower the cost per target when compared with shotgun sequencing.

In the present study, we describe the first adaptation of the SHS capture method for the selective enrichment of a target-specific biomarker from a complex environmental metagenome. Methane (CH_4) is an important radiative trace gas responsible for the greenhouse effect, and a significant proportion (6–16%) of the global natural methane emissions are released from freshwater lakes.[25] We surveyed the methanogen diversity in a permanently stratified crater lake located in the French Massif Central (Lake Pavin). This original freshwater ecosystem is composed of an anoxic deep water layer (monimolimnion, ~60–90 m depth) separated from the oxygenated upper layer (mixolimnion) by an intermediate layer (mesolimnion),[26] where both the sediments and the anoxic water column contribute to methane production.[27] We targeted the gene coding for the α-subunit of the methyl coenzyme M reductase (*mcrA*) that is involved in the final step of methanogenesis. This gene is arranged in a single transcriptional unit—the *mcr* operon—that is highly conserved among all methanogens.[28,29] To highlight the broad benefits of the gene capture approach when compared with the more classical sequencing methods, three methods were used for pyrosequencing of an environmental sample: the SHS method, a classical random-shotgun metagenomic approach and an *mcrA*-targeted amplicon sequencing survey.

2. Materials and methods

2.1. Capture probe design and synthesis

Two sets of capture probes were designed. The first set targeted the *mcrA* gene from the *Methanosarcina acetivorans* C2A genome (GenBank accession no. AE010299), and the second set targeted the *mcrA* sequences pooled from environmental samples. The first set of capture probes consisted of six high specific 50-mer probes (P1–P6) targeting six distinct regions of the *M. acetivorans* C2A *mcrA* gene (Fig. 2, Supplementary Table S1). These probes were designed with HiSpOD software.[30] Adaptor sequences were added at each end, resulting in 80-mer hybrid probes consisting of 5′-ATCGCACCAGCGTGT$(X)_{50}$C ACTGCGGCTCCTCA-3′, with X_{50} indicating the specific capture probe.

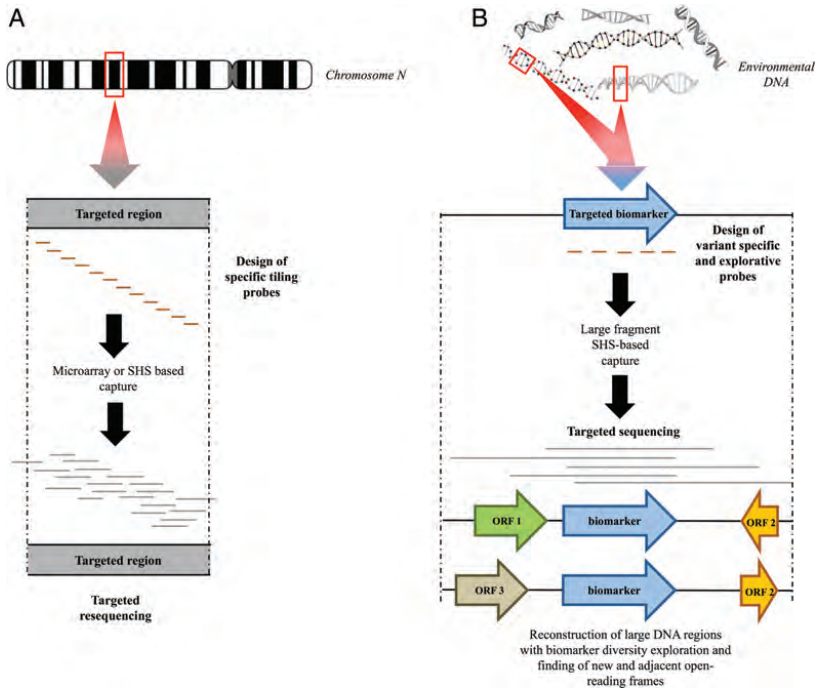


Figure 1. Schematic comparison of targeted capture methods applied to classical direct selection method of individual genomic loci (human for instance) (A) and our new approach for metagenomics targeting (B). The enrichment through microarray and the SHS of large genomic regions within complex eukaryotic genomes, as described in A, uses specific tiling probes to target resequencing genomic loci for copy number variation (CNV) and single nucleotide polymorphism detection. Our SHS method (B) uses the design of specific variants and explorative probes across a targeted biomarker to specifically enrich large DNA fragments from complex metagenomic DNA. Captured DNA fragments are sequenced to explore biomarker diversity and adjacent flanking regions. The red rectangles indicate the targeted regions.

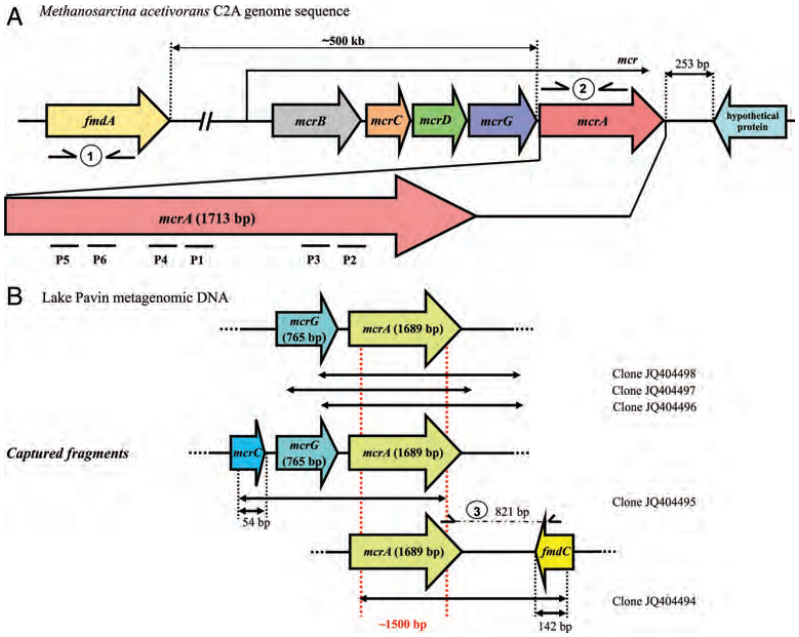


Figure 2. Schematic representation of *mcr* operon fragments on (A) *M. acetivorans* C2A gDNA and (B) Lake Pavin metagenomic DNA. Primer pairs used for *fmdA* (1) and *mcrA* (2) quantification as well as *mcrA-fmdC* region (3) amplification are symbolized. Dashed arrows indicate the sequence coverage of each of the five clones retrieved from the environmental sample (B). P1–P6: Positions of the six capture probes in the *mcrA* gene of *M. acetivorans* (see Supplementary Table S1 for probe sequences).

The second set of capture probes was 26 oligos (1 49-mer and 25 50-mers) designed to target *mcrA* and *mrtA* (encoding the α-subunit of the methyl coenzyme M reductase isoform II), but not the *mcrA* of anaerobic methanotrophs (Supplementary Table S2, Supplementary Fig. S1).

Oligonucleotides were purchased from Eurogentec S.A. (Belgium). The RNA probe was prepared as described by Gnirke *et al.*[19]

2.2. Preparation of biological samples and libraries

The two biological models used in this study were the *M. acetivorans* C2A strain (DSM 2834) and Lake Pavin, located in the French Massif Central (45°29′74″N, 2°53′28″E). The *M. acetivorans* C2A strain was cultivated using the medium 304 (http://www.dsmz.de/microorganisms/medium/pdf/DSMZ_Medium304.pdf) according to the manufacturer's instructions. gDNA from the strain was extracted using the Easy DNA kit (Invitrogen), whereas environmental DNA was extracted from 350 ml of freshwater collected from Lake Pavin at a 90-m depth, as described by Dugat-Bony *et al.*[30]

Libraries were prepared using Roche's GS FLX Titanium General Library Preparation Kit (Roche Applied Science) according to the manufacturer's instructions. First, 5 μg of DNA was sheared by nebulization. DNA fragments were size selected with AMPure beads (Beckman Coulter Genomics). After purification, fragment end polishing, adaptor ligation (A and B adapter keys; Supplementary Table S1) and fill-in reactions, the libraries were PCR amplified with the 454 Ti-A and 454 Ti-B primers (Supplementary Table S1). The cycle conditions were 3 min at 93°C followed by 20 cycles of 15 s at 93°C, 1 min at 58°C and 8 min at 68°C and a final elongation step at 68°C for 6 min. The amplified libraries were purified with AMPure beads and stored at −20°C until use.

For the amplicon library, *mcrA* fragments were PCR amplified from total community DNA with the *mcrA*-specific primer pair MM_01/MM_02[31] (Supplementary Table S1). The amplicon was run on a 2% (wt/vol) agarose gel, and the ∼500 bp product was purified with a QIAquick gel extraction kit (Qiagen) and AMPure beads. Each DNA library was quantified by fluorometry with a Quant-iT PicoGreen dsDNA assay kit (Invitrogen). The DNA quality and size distribution were assessed on an Agilent Bioanalyzer High Sensitivity DNA chip (Agilent Technologies).

2.3. Hybridization capture and elution

For each SHS-capture method library, 2.5 μg of salmon sperm DNA (Ambion) and 500 ng of DNA library were mixed (7 μl final volume), denatured for 5 min at 95°C, incubated for 5 min at 65°C before adding 13 μl of prewarmed (65°C) hybridization buffer (10X SSPE, 10X Denhardt's Solution, 10 mM EDTA and 0.2% SDS) and 6 μl freshly prepared, prewarmed (2 min at 65°C) biotinylated RNA probes (500 ng). After 24 h at 65°C, 500 ng of washed M-280 Dynabeads coated with streptavidin (Invitrogen) were added to the hybridization mix that was incubated for 30 min at room temperature (RT). The beads were precipitated with a magnetic stand (Ambion) and washed once for 15 min at RT with 500 μl 1X SSC/0.1% SDS and three times for 10 min at 65°C with 500 μl prewarmed 0.1X SSC/0.1% SDS. The captured DNA was eluted with 50 μl 0.1 M NaOH for 10 min at RT. After magnetic bead precipitation, the DNA supernatant was transferred to a sterile tube containing 70 μl of 1 M Tris–HCl pH 7.5, purified on a QIAquick column (Qiagen) and eluted in a final volume of 20 μl. A 2.5 μl aliquot was subjected to 15 cycles of PCR amplification using the 454 Ti-A and Ti-B primers as described above. After purification, a second round of capture was performed from each first-round PCR product. To increase the DNA yield, a final PCR amplification consisting of 20 cycles was performed. The final product was purified on a QIAquick column (Qiagen) and quantified with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies).

2.4. Sanger sequencing and data analysis

PCR products were cloned using the TOPO TA cloning kit (Invitrogen). Plasmids were screened for high-size inserts by digestion with *Eco*RI, and positive clones were Sanger sequenced at MWG DNA sequencing services (Ebersberg, Germany). Sequences were processed and joined using the Staden package program,[32] and primer sequences were removed from paired-end consensus sequences. The *mcr* sequence data retrieved from Lake Pavin by the SHS method were deposited in the GenBank database under accession numbers JQ404494, JQ404495, JQ404496, JQ404497 and JQ404498, and the sequence of the *mcrA-fmd* region-spanning fragment was deposited under accession number JQ425691.

2.5. 454 GS FLX Titanium sequencing and data analysis

DNA samples were sequenced using the GS FLX Titanium system on the 'GINA' platform (part of the GENTYANE platform, labelled IBISA since 2009; BP 392, 63 011 Clermont-Ferrand, France) at the Centre Jean Perrin, according to the manufacturer's specifications. For quality filtering and de-replication of reads, sequences were trimmed with the PRINSEQ-lite PERL script[33] using the parameters described in

the preprocessing chart (http://prinseq.sourceforge.net/Preprocessing_454_SFF_chart.pdf).

Functional assignment and enrichment were assessed with a BLASTX query[34] against a database containing 12 603 McrA protein sequences downloaded from the Genbank database (http://www.ncbi.nlm.nih.gov/), using WWW-Query (http://pbil.univ-lyon1.fr/search/query_fam.php) to perform an advanced keyword search. Reads showing >40% identity over 100 or more amino acids were classified as McrA sequences. Chimaera detection was performed with the UCHIME program[35] with a stringent threshold score of five. Sequences containing possible frameshifts were identified with the '−w 20' BLAST option and disabled low complexity filters. Amino acid sequences without frameshifts were extracted from the BLAST results, and only the sequences that passed this filter were chosen for further phylogenetic analysis.

The sequence data were deposited in the NCBI as a Short Read Archive (SRA) project under accession no. SRA049219.

2.6. *Phylogenetic analysis and tree construction*

All McrA sequences derived from the SHS method and metagenomic libraries were aligned to a sequence obtained from the amplicon library. The amino acid alignment used the ClustalW2 alignment method[36] driven by the Seaview version 4 program[37] to select the reads having at least 100 amino acids in common with this reference sequence. The overlapping regions of the remaining amino acid sequences, all amplicon pyrosequences and 29 McrA sequences previously identified from the same sampling depth and downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) were fed to CD-HIT[38] that assigned them to operational taxonomic units (OTUs) using a complete linkage clustering method at a 91% cut-off value.[27,39]

One representative sequence of each OTU was chosen to build a phylogenetic tree (Seaview 4)[37] using the neighbour-joining method[40,41] and 1000 bootstrapped trials. Closely related sequences available from GenBank (http://www.ncbi.nlm.nih.gov/) were included in the phylogenetic trees to decipher the microbial community diversity. A final tree was drawn in MEGA version 5.[42]

2.7. *qPCR assays for enrichment and methanogen abundance*

The assays were conducted in 20 µl with 5 µl of DNA sample or *mcrA* PCR product standards (covering a dynamic range of 5×10^7 to 50 copies), 10 µl of 2X MESA Green quantitative Polymerase Chain Reaction (qPCR) for SYBR assay mixture (Eurogentec S.A) and 0.2 µM forward and reverse primers. The thermo cycling protocol included an initial step of 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 15 s, annealing at the melting temperature of each primer set for 15 s and elongation at 68°C for 30 s. The samples and each point of the standard curve were quantified in triplicate. The primer sets are described in Supplementary Table S3. The data were analysed with Realplex software version 1.5 (Eppendorf Inc.) and MxPro qPCR software 4.10d (Agilent technologies). Based on the ΔΔCt method,[43] relative enrichments (R) were calculated according to $R = 2^{-\Delta\Delta Ct}$. The relative quantification method established a mean Ct value comparison (ΔCt) between *mcrA* (target gene) and *fmdA* (non-target gene 500 kb upstream from *mcrA*). The relative capture enrichment was determined by the comparison of ΔCt before and after capture, and this result described the fold change or ΔΔCt.

2.8. *SHS* de novo *read assembly*

The filtered SHS reads were assembled with Newbler version 2.6 (Roche Applied Science) using stringent assembly parameters (60 bases overlap and 95% overlap identity) and the '- rip' option that forces Newbler to place each read into one unique contig. The functional assignment of contigs and singletons was performed by a BLASTX query[34] against our database containing 12 603 McrA protein sequences. Chimaeras were detected in the *mcrA* contigs and singletons with the UCHIME program[35] and a stringent threshold score of five. Prediction of the *mcrA* gene location within contigs and singletons was performed by BLASTN[44] against the reference genomes of *Candidatus Methanoregula boonei* 6A8 (*Methanomicrobiales* order, accession no. NC_009712), *Methanosaeta concilii* GP-6 (*Methanosarcinales* order, accession no. CP002565) and *Methanosphaera stadtmanae* DSM 3091 (*Methanobacteriales* order, accession no. CP000102). Contigs extending at least 100 nucleotides beyond *mcrA* were segregated for BLASTX[34] analysis against the non-redundant (nr) protein sequences database to identify putative open-reading frames within the flanking regions.

The sequence data from homologous *mcrA* contigs (without chimaeras or frameshifts) were deposited in the GenBank database under accession no. KC184908 to KC185399.

3. Results

3.1. *Development of an SHS method for genomic-scale sequence enrichment*

3.1.1. *Method validation: mcrA gene enrichment from* M. acetivorans *C2A gDNA*

We performed the initial validation of our enrichment

strategy by capturing the *mcrA* gene from a 1 to 3 kb fragment library of the completely sequenced methanogenic *M. acetivorans* C2A strain. The minimal probe set spanned different non-overlapping regions of the gene (Fig. 2A). The qPCR reactions revealed a 461-fold relative enrichment of *mcrA* sequences after the first cycle of capture and at least 175 365-fold enrichment after the second cycle. Furthermore, as the *M. acetivorans* C2A genome consists of 5751 kb with a single *mcrA* gene copy, the probability of randomly sequencing this gene from a 1 to 3 kb fragment size clone library is 0.02–0.05%. Using our solution-based DNA capture-enrichment method and working on an isolated species, the likelihood increased from 7.8 to 23% after the first cycle and could reach 100% after the second.

The DNA sequence of fragments retrieved after the second cycle of capture was controlled by the cloning-sequencing method. Six clones were sequenced, and all had a perfect correspondence to the *mcrA* gene from *M. acetivorans* C2A, reinforcing the efficiency of the two iterative cycles of capture. The captured fragments were assembled into a 1834-bp contig containing the nearly complete *mcrA* gene (1645 bp) and its 3′ non-coding region (189 bp). After validating this approach, we further tested the performance of the method by enriching *mcrA* sequences from a complex methanogenic freshwater environment.

3.1.2. Environmental application: mcrA sequence enrichment from a methanogenic lacustrine environment (Lake Pavin) The freshwater sample was collected in the anoxic zone at 90 m depth, where the highest methanogen diversity was available in the lacustrine environment.[27] An improved *mcrA* probe set included all known *mcrA* sequences and targeted new variants with explorative probes (Supplementary Table S2). The efficiency of the *mcrA* enrichment was determined by cloning and sequencing the second capture product. Five out of the ten clones with large inserts (2041–2493 bp) included *mcrA* sequences. All positive clones had a ~1500 bp common zone corresponding to the *mcrA* gene, but they also harboured upstream or downstream regions containing other genes (Fig. 2B). BLAST analysis of the cloned sequences revealed that they are very similar (99% similarity) to *mcrA* sequences previously retrieved from this ecosystem (accession nos. GQ389949, GQ389912 and GQ389806).[27] The closest relative to the *mcrA*, *mcrG* and partial *mcrC* sequences were from a cultured methanogen, *Candidatus Methanoregula boonei* 6A8 (>85, 84 and 81% similarity, respectively). This hydrogenotrophic species belongs to the *Methanomicrobiales* order, and it was isolated from an acidic peat bog.[45] Furthermore, the *fmdC* gene fragment identified 821 bp downstream the target gene (Fig. 2B) that shared 77% identity with subunit C of the formyl methanofuran dehydrogenase gene of this species. This gene has been located in the reference genome (GenBank: CP000780.1) at almost 300 kb from the *mcr* operon. It should be noted that this genome organization—with the *fmd* operon located just downstream from the *mcr* operon—has not been described previously in methanogens. To exclude the possibility of chimaera formation during metagenomic library amplification, a PCR fragment spanning the *mcrA–fmdC* region was obtained directly from the initial metagenomic DNA sample, using two specific primers (Fig. 2B, Supplementary Table S1). The sequenced 821 bp PCR product (JQ425691) confirmed the organization revealed by the SHS method (100% identity with the captured DNA fragment).

Our results showed that the capture method not only efficiently enriched targets out of a complex environmental genomic mixture, but also recovered sequences adjacent to the targeted biomarker gene. Additionally, the SHS method was coupled with NGS technologies to assess the coverage of archaeal *mcrA* diversity in a complex ecosystem.

3.2. Metagenome exploration with genome-scale sequence enrichment and NGS

The benefit of the SHS method in terms of diversity coverage, when compared with more classical approaches, was further examined by sequencing the SHS capture products. A new random-shotgun DNA metagenomic library adapted for pyrosequencing (fragment sizes ~500 bp) was prepared for the SHS products and for direct sequencing (shotgun metagenomics approach). From the same metagenomic DNA sample, *mcrA* PCR products were also amplified with the primer set MM_01-MM_02,[31] generating amplicons of ~500 bp. Sequencing (captured DNA fragments, metagenome and amplicons) was performed with the 454 GS FLX Titanium technology, generating a slightly different amount of raw data with an average read length of 414–471 bases. After pre-processing, sequencing datasets from all three approaches had nearly equivalent numbers of reads (Table 1).

3.2.1. Functional assignment and enrichment performance Only three reads (0.003% of total reads) from the random-shotgun sequencing approach corresponded to the *mcrA* gene. For the SHS method, 50 727 reads were identified as *mcrA* sequences (41.32%), and almost all the amplicon approach sequences were from *mcrA* (119 409 reads, 99.98%).

For *mcrA* diversity evaluation, however, we only analysed high-quality sequences (no chimaeras or

Table 1. Summary statistics from 454 pyrosequencing

	Metagenome	Amplicons	SHS
Total number of raw reads	136 256	121 665	177 977
Number of reads after pre-processing	116 365	119 437	122 772
Average length of cleaned reads (bases)	471	414	454
mcrA homologous sequences[a]	3	119 409	50 727
Enrichment performance (%)	0.003	99.98	41.32
Number of chimaeras	0	150	30
Number of reads containing frameshifts	1	80 390	21 855
Number of high-quality *mcrA* homologous sequences (without chimaera and frameshifts)	2	38 869	28 842
McrA sequences used for methanogenic diversity and abundance (comparison of a common region)	1	38 807	11 442
Number of OTUs	1	40	44
McrA sequences related to OTUs	1	38 784[b]	11 324[b]
Relative abundance of *mcrA* sequences affiliated with *Methanomicrobiales* (%)	0	98.57	98.82
Relative abundance of *mcrA* sequences affiliated with *Methanosarcinales* (%)	0	0.005	0.86
Relative abundance of *mcrA* sequences affiliated with the Novel Order (%)	100	1.43	0.13
Relative abundance of *mcrA* sequences affiliated with *Methanobacteriales* (%)	0	0	0.19

[a]BLASTX parameters: percentage of identity: 40%; E-value cut-off: 10.
[b]McrA sequences related to OTUs containing more than one sequence.

frameshifts), and all the problematic reads were subsequently excluded.

3.2.2. Methanogen diversity and abundance The phylogeny of the methanogenic McrA protein sequences was investigated and compared for each of the three approaches. We used ClustalW2[36] to determine a common reference region of 143 amino acids shared by the largest number of McrA sequences retrieved from the 3 approaches. All McrA sequences that included this region were truncated so that at least 100 amino acids aligned with this reference. The resulting sequences, which included 1 read from the shotgun library, 11 442 reads from the SHS method library and 38 807 reads from the amplicon library, were used for further analysis. Furthermore, 29 additional sequences (referred to as Pavin90m) from a previous study[27] were included in the analysis.

Following the clustering method, 127 distinct OTUs (longer than 300 bp) were observed, and the 58 OTUs that contained more than 1 sequence were included in a more detailed phylogenetic analysis. The shotgun library sequence, which contained a single final read, was also included. Among these 58 OTUs, 44 were detected from the SHS method, 40 from the amplicon approach, 1 from the metagenomic shotgun library and 3 from Pavin90m sequences. The SHS method and amplicons shared 27 OTUs, including 3 from the Pavin90m sequences (Fig. 3). The remaining 31 OTUs were specific to a single method, with 1 for the metagenome, 17 for the SHS and 13 for the amplicons (Fig. 3).

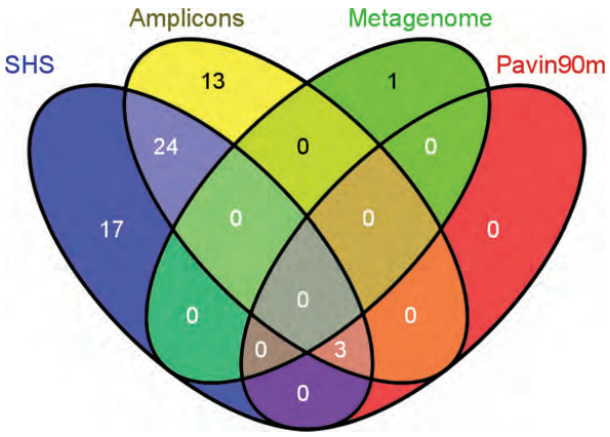


Figure 3. Venn diagram showing the number of unique and shared OTUs for the in-solution capture method (SHS), PCR-based strategy (Amplicons) and sequences isolated at 90 m depth from a previous PCR-based study of Lake Pavin (Pavin90m).[27] The Venn diagram was generated with Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

The 58 OTUs covered four lineages including *Methanobacteriales*, *Methanomicrobiales*, *Methanosarcinales* and a putative fourth lineage called 'Novel Order'. Most OTUs were closely related to the *Methanomicrobiales* order (48 OTUs, 98.6% of the total input sequences). OTU3, OTU10 and OTU17 formed a distinct branch within this cluster (Fig. 4A), and they were closely related to cultured methanogenic species that also have an insertion in their McrA protein sequence (Supplementary Fig. S2). Both the SHS and amplicon

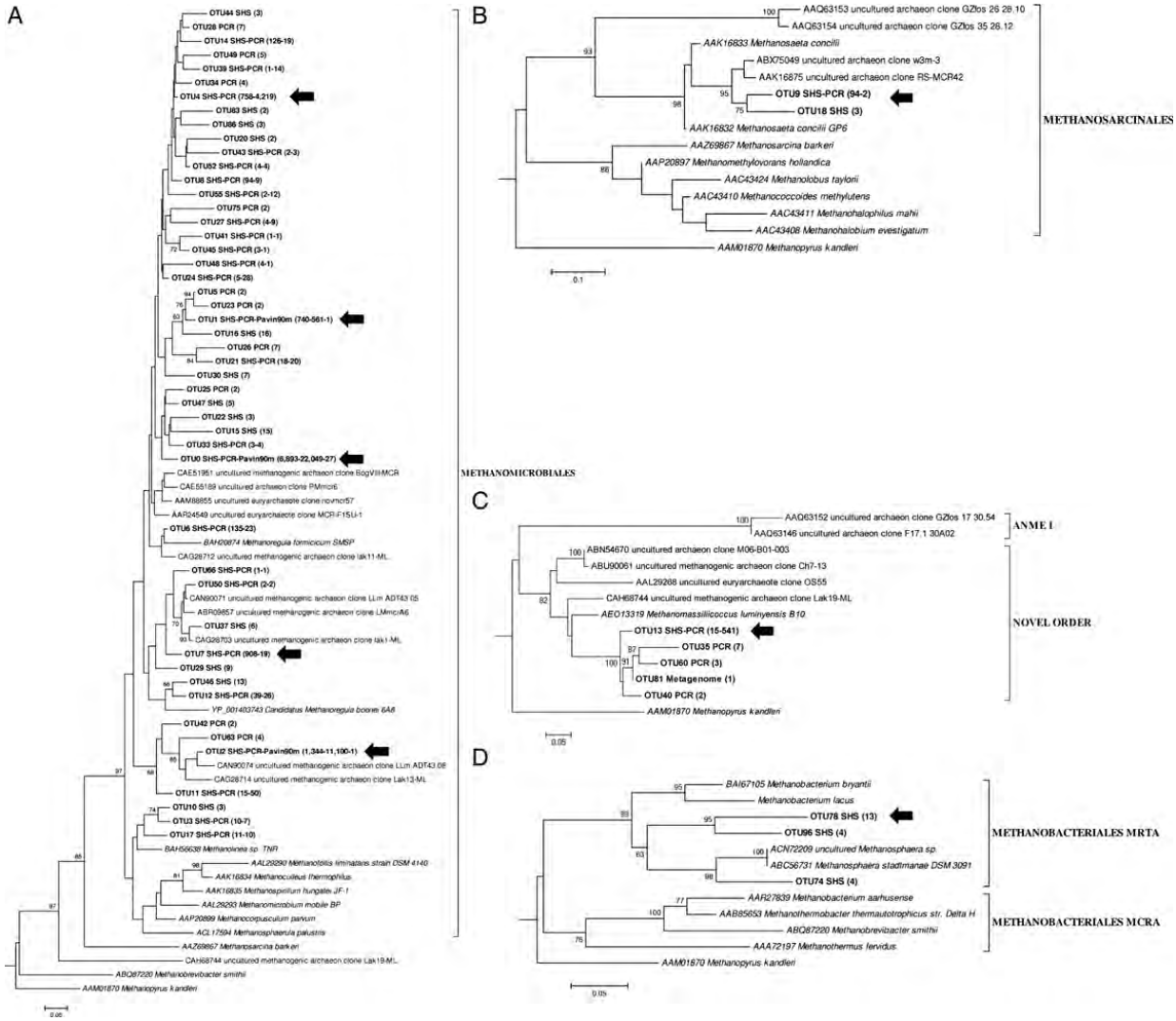
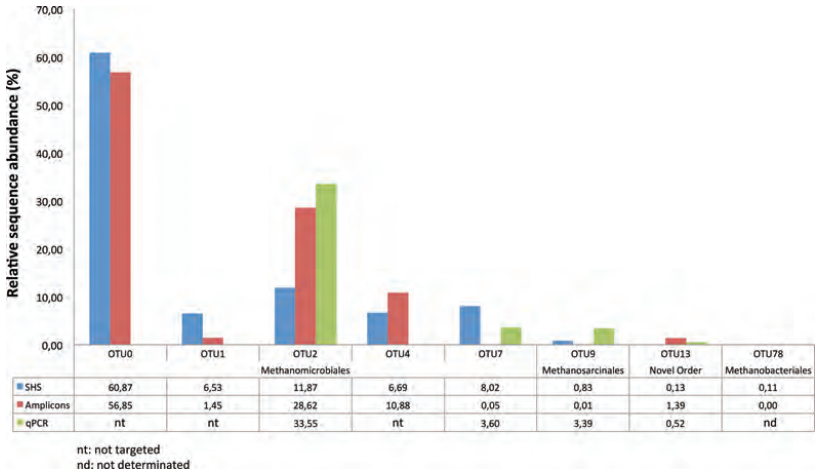


Figure 4. Phylogenetic analysis of deduced McrA amino acid sequences obtained from the PCR, SHS and Pavin90m datasets showing evolutionary distances within the orders *Methanomicrobiales* (A), *Methanosarcinales* (B), Novel Order (C) and *Methanobacteriales* (D). Evolutionary history was inferred using the neighbour-joining method[40,41] (Poisson distance model) using Seaview software.[37] The final tree was drawn in MEGA 5.[42] The bars represent a 5% sequence divergence. Numbers at the nodes represent bootstrap values >60% (1000 resamplings). The number of amino acid sequences assigned to each OTU is given in brackets, together with the name of the strategies for obtaining them. McrA amino acid sequence from *Methanosarcina barkeri* (AAZ69867), uncultured methanogenic archaeon clone Lak19-ML (CAH68744) and *Methanobrevibacter smithii* (ABQ87220) were used as outgroups, and *Methanopyrus kandleri* (AAM01870) was an outgroup for rooting the tree. Bold arrows indicate dominant OTUs.

strategies clustered sequences in the most abundant OTUs (Fig. 5). These abundant OTUs represented 94 and 98%, respectively, of the total sequences for each approach. The *Methanosarcinales* (two OTUs; Fig. 4B) grouped into two distinct branches were related to the reference acetoclastic species *M. concilii* GP6 (85 and 87% similarity with OTU9 and OTU18, respectively). The most abundant cluster was OTU9 that represented 0.83% of the total SHS reads and 0.005% for the total amplicon reads (Fig. 5). In contrast, the putative Novel Order (five OTUs; Fig. 4C) was dominated by OTU13 clustering with 1.39% of the total amplicons sequences, but only 0.13% of the total SHS reads (Fig. 5). Even if we did not include the more recently described sequences of *Methanomassiliicoccus luminyensis*[46] and *Candidatus Methanomethylophilus alvus*[47] belonging to the novel order for the probe design, distant sequences could be captured with probes by a mismatched nucleotide pairing. We cannot exclude that the sequences captured by specific probes allow indirect hybridization of other *mcr*A sequences as described for DNA microarrays experiments and referred to as 'hitchhiking'.[48] Despite the substantial sequencing effort for amplicons, no



	OTU0	OTU1	OTU2 Methanomicrobiales	OTU4	OTU7	OTU9 Methanosarcinales	OTU13 Novel Order	OTU78 Methanobacteriales
SHS	60,87	6,53	11,87	6,69	8,02	0,83	0,13	0,11
Amplicons	56,85	1,45	28,62	10,88	0,05	0,01	1,39	0,00
qPCR	nt	nt	33,55	nt	3,60	3,39	0,52	nd

Figure 5. The relative abundances of dominant OTUs from four methanogenic bacterial orders identified by the targeted capture method (SHS), PCR-based strategy (amplicons) and qPCR experiments (qPCR). The relative abundances calculated by qPCR were computed using *mcrA* copy number as reference obtained using a primers pair targeting all OTUs (Supplementary Table S3).

sequences belonging to the *Methanobacteriales* order were recovered from this approach. These sequences were obtained only from the SHS sample (Fig. 4D), and they were clustered in three OTUs such that one was 90% similar to MrtA (MCR isoenzyme encoded by the *mrt* operon) from *M. stadtmanae* DSM 3091[49] and the remaining two were 77 and 79% identical to MrtA sequences from *Methanobacterium lacus* that is in the *Methanobacteriales* order and has been isolated from Lake Pavin sediments.[50] These sequences represented 0.19% of total SHS *mcrA*-related sequences, with the most abundant OTU78 clustering 0.11% of the total SHS reads (Fig. 5).

The GC content of the *mcrA* genes ranged from 50.4 to 61.1% for amplicons and from 37 to 63.2% for SHS. In the *mcrA* database, the GC content ranges from 36.2 to 67.2%, indicating that the SHS method is most likely less affected by GC composition than PCR approaches. Furthermore, we evaluated the presence of mismatch residues between PCR primers and probes on *mcrA* genes in both SHS and amplicon approaches. We identified 99.10, 0.77 and 0.13% of *mcrA* sequences for amplicons versus 37.68, 50.22 and 12.10% for SHS with 0, 1 and 2 mismatch residues, respectively, between probes (or primers) and sequences. This trend highlights the potential advantage of the SHS approach with long capture probes that tolerate more mismatches, allowing access to new *mcrA* gene variants.

In parallel, qPCR was used to precisely describe the methanogen abundance in Lake Pavin with regard to the most abundant OTUs and bacterial orders (primers are listed in Supplementary Table S3). The results were compared with the relative sequence abundance calculated previously for the selected OTUs with amplicons and SHS (Fig. 5). The abundance of OTU2, which included the *Methanomicrobiales* order, was similar in qPCR and amplicons (33.5 and 28.62%), but not SHS (11.87%). In contrast, the second *Methanomicrobiales* OTU (OTU7) was more abundant in SHS (8.02%) and qPCR (3.6%), but not amplicons (0.05%). The same trend was observed for OTU9 (*Methanosarcinales*). No significant difference was observed for OTU13 (Novel Order). Finally, no qPCR amplification of OTU78 (*Methanobacteriales*) occurred. However, we validated the presence of this OTU in Lake Pavin by successive PCR cycles, cloning and sequencing (100% identity). This result indicates that *Methanobacteriales* are rare in this ecosystem.

3.2.3. De novo assembly of SHS reads To reconstruct contigs with sequences flanking the targeted *mcrA* gene, *de novo* assembly was performed using the pyrosequencing reads obtained by the SHS method (Table 2). We identified 691 contigs (301–1639 bases) with *mcrA* sequences. By mapping these sequences to complete reference genomes for the *Methanomicrobiales*, *Methanosarcinales* and *Methanobacteriales* orders (no genome was available for the Novel Order), we identified contigs extending into the *mcrA* flanking regions. The upstream sequences were all part of the *mcrG* gene. We also characterized two adjacent ORFs located at 200 bases downstream from the *mcrA* gene and in the same orientation; these ORFs encoded a DtxR family iron (metal)-dependent repressor and a DOMON domain-containing protein. The DtxR sequences were closely related (76–83% identity) to *Methanosphaerula palustris* E1-9C (accession no. ACL16981) of the *Methanomicrobiales* order. In the reference genome of this species, the gene is located ~700 kb downstream of the *mcr* operon. The

Table 2. Summary statistics from *de novo* assembly

Newbler version 2.6	SHS
No. of reads used for assembly	122 772
No. of reads assembled into contigs	53 307
No. of singletons	56 834
Outliers[a]	12 631
No. of contigs assembled	1916
N_{50} contig size (bases)	820
No. of *mcrA* homologous contigs	693
No. of *mcrA* homologous singletons	1142
Number of chimaeras	5
Number of high-quality *mcrA* homologous contigs (without chimaeras)	691
Number of high-quality *mcrA* homologous singletons (without chimaeras)	1139
Average *mcrA* homologous contig length (bases)	589
Largest *mcrA* homologous contig length (bases)	1639

[a]Reads were discarded due to quality control by Newbler.

sequences of DOMON domain-containing protein are closely related (74–80% identity) to *M. concilii* GP-6 (accession no. AEB67518) that belongs to the *Methanosarcinales* order. In the reference genome of this species, the gene is located ~50 kb downstream of the *mcr* operon.

4. Discussion

We captured specific target DNA from a complex environmental metagenome using a novel SHS capture method and NGS. We showed that the relative enrichment of the target sequence was increased to 175 365-fold with 2 cycles of capture, and this result was superior to previous studies using a single cycle[18,19] and microarray-based capture.[51] We applied this strategy to the anoxic layer of Lake Pavin, where *Archaea* account for 17% of 4,6-diamidino-2-phenylindole-stained cells[52] and only a fraction of these microbes are methanogens. Our SHS strategy specifically enriched *mcrA* sequences from the environmental sample. In comparison with the random-shotgun metagenomic approach (0.003% recovery of *mcrA* sequences), the SHS method was superior (41.32% *mcrA* sequence enrichment). However, the capture efficiency is also likely influenced by the number of probes used per region and the mismatched residues between the probes and their targets. Consequently, two rounds of capture and multiple long RNA probes are advantageous for efficient enrichment.

With a random-shotgun metagenomics approach, many hundreds of thousands of additional single reads would have been necessary to estimate the biodiversity of the methanogen community in this environment. The SHS experiment contained much more *mcrA* data and provided a solid taxonomic basis for studying methanogens diversity. Finally, PCR was the most effective enrichment approach; with ~100% of the amplicons corresponding to the biomarker, the primers used were very specific and efficient.[31]

The SHS and amplicon strategies both revealed similar patterns in methanogen communities such as the high abundance and diversity of *Methanomicrobiales* sequences (more than 98% of the total sequences representing 48 OTUs). These data confirm a previous study by Biderre-Petit *et al.*[27] High-throughput sequencing, however, reveals that methanogen diversity is much higher than previously estimated by amplicon libraries and Sanger sequencing.[27] Importantly, the amplicon sequencing approach missed all the *Methanobacteriales* taxonomic groups and some *Methanosarcinales*, possibly due to *mcrA* primer bias. PCR undersampling often leads to significant underestimation of true community diversity.[24,53] SHS efficiently targets rare sequences, as demonstrated for *Methanobacteriales*, and does not appear to be influenced by GC content. As previously demonstrated for microarray approaches,[21,22,54] more extensive explorative capture probe sets could recover rare sequences, leading to the detection of many uncharacterized microbial populations. Moreover, the SHS and amplicon library results were correlated by qPCR.

We also used *de novo* assembly of SHS sequence reads to explore the regions flanking the target gene, and we identified two ORFs (*dtxR* and DOMON domain) at previously unknown positions downstream of *mcrA*. Because this genomic organization may link methanogenesis to electron transfer and Fe homeostasis in organisms living in the anoxic layer of the Lake Pavin, it could reflect adaptation to this particular environment. More experiments are needed, however, to validate this hypothesis.

In this study, we present a novel enrichment method that, when coupled to NGS, expands our knowledge of the diversity of a target gene within a complex microbial community. The method was successfully applied to a lacustrine environment using the *mcrA* gene, and it revealed higher methanogen community diversity than observed with other methods. To some extent, this method could be applied to phylogenetic studies to explore the diversity of commonly conserved genes such as the 16S rRNA biomarker. The main limitation is the design of high quality probes sets to expect a full coverage of 16S rDNA sequences as complete as possible. New algorithms, such as KASpOD,[55] can be used to design highly specific and explorative probes (i.e. targeting sequences not already included in databases) based on oligonucleotide *k-mer* signatures. These probe designs would be extremely suitable and beneficial to the SHS approach.

With the emergence of third generation sequencing platforms and the capability to sequence longer DNA sequences without library construction,[56,57] the SHS strategy could link genomic structure and function in microbial communities.

Acknowledgements: We would like to thank Yannick Bidet and Maud Privat from the Centre Jean Perrin for their help with sample processing on the 454 GS FLX pyrosequencing platform. We also thank Sarah Orlhac and Nicolas Gallois for their efficient technical assistance and David Tottey for reviewing the English version of the manuscript.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the ANR-09-EBIO-009 project (Agence Nationale de la Recherche). J.D. was supported by a studentship from the Centre National de la Recherche Scientifique (CNRS, grant number 163588) and the Région Auvergne. N.P. was funded by Direction Générale de l'Armement (DGA).

References

1. Whitman, W.B., Coleman, D.C. and Wiebe, W.J. 1998, Prokaryotes: the unseen majority, *Proc. Natl. Acad. Sci. USA*, **95**, 6578–83.
2. Curtis, T.P., Head, I.M., Lunn, M., Woodcock, S., Schloss, P.D. and Sloan, W.T. 2006, What is the extent of prokaryotic diversity? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **361**, 2023–37.
3. Amann, R., Ludwig, W. and Schleifer, K.-H. 1995, Phylogenetic identification and in situ detection of individual microbial cells without cultivation, *Microbiol. Rev.*, **59**, 143–69.
4. Eisen, J.A. 2007, Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes, *PLoS Biol.*, **5**, e82.
5. Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J. and Goodman, R.M. 1998, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chem. Biol.*, **5**, R245–249.
6. Biddle, J.F., Fitz-Gibbon, S., Schuster, S.C., Brenchley, J.E. and House, C.H. 2008, Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment, *Proc. Natl. Acad. Sci. USA*, **105**, 10583–88.
7. Tringe, S.G., von Mering, C., Kobayashi, A., et al. 2005, Comparative metagenomics of microbial communities, *Science*, **308**, 554–7.
8. Riesenfeld, C.S., Schloss, P.D. and Handelsman, J. 2004, Metagenomics: genomic analysis of microbial communities, *Annu. Rev. Genet.*, **38**, 525–52.
9. Suenaga, H. 2011, Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities, *Environ. Microbiol.*, **14**, 13–22.
10. Edwards, R.A., Rodriguez-Brito, B., Wegley, L., et al. 2006, Using pyrosequencing to shed light on deep mine microbial ecology, *BMC Genomics*, **7**, 57.
11. Mardis, E.R. 2008, The impact of next-generation sequencing technology on genetics, *Trends Genet.*, **24**, 133–41.
12. Quince, C., Curtis, T.P. and Sloan, W.T. 2008, The rational exploration of microbial diversity, *ISME J.*, **2**, 997–1006.
13. Hoff, K.J. 2009, The effect of sequencing errors on metagenomic gene prediction, *BMC Genomics*, **10**, 520.
14. Summerer, D. 2009, Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing, *Genomics*, **94**, 363–8.
15. Albert, T.J., Molla, M.N., Muzny, D.M., et al. 2007, Direct selection of human genomic loci by microarray hybridization, *Nat. Methods*, **4**, 903–5.
16. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. 2007, Microarray-based genomic selection for high-throughput resequencing, *Nat. Methods*, **4**, 907–9.
17. Mokry, M., Feitsma, H., Nijman, I.J., et al. 2010, Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries, *Nucleic Acids Res.*, **38**, e116.
18. Tewhey, R., Nakano, M., Wang, X., et al. 2009, Enrichment of sequencing targets from the human genome by solution hybridization, *Genome Biol.*, **10**, R116.
19. Gnirke, A., Melnikov, A., Maguire, J., et al. 2009, Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing, *Nat. Biotechnol.*, **27**, 182–9.
20. Iwai, S., Chai, B., Sul, W.J., Cole, J.R., Hashsham, S.A. and Tiedje, J.M. 2010, Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment, *ISME J.*, **4**, 279–85.
21. Terrat, S., Peyretaillade, E., Goncalves, O., et al. 2010, Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development, *BMC Bioinformatics*, **11**, 478.
22. Dugat-Bony, E., Peyretaillade, E., Parisot, N., et al. 2011, Detecting unknown sequences with DNA microarrays: explorative probe design strategies, *Environ. Microbiol.*, **14**, 356–371.
23. Suzuki, M. and Giovannoni, S. 1996, Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR, *Appl. Environ. Microbiol.*, **62**, 625–30.
24. Hong, S., Bunge, J., Leslin, C., Jeon, S. and Epstein, S.S. 2009, Polymerase chain reaction primers miss half of rRNA microbial diversity, *ISME J.*, **3**, 1365–73.
25. Bastviken, D., Cole, J., Pace, M. and Tranvik, L. 2004, Methane emissions from lakes: dependence of lake characteristics, two regional assessments, and a global estimate, *Global Biogeochem. Cycles*, **18**, GB4009, doi:10.1029/2004GB002238.

26. Aeschbach-Hertig, W., Hofer, M., Kipfer, R., Imboden, D.M. and Wieler, R., 1999, Accumulation of mantle gases in a permanently stratified volcanic lake (Lake Pavin, France), *Geochim. Cosmochim. Acta*, **63**, 3357–72.
27. Biderre-Petit, C., Jezequel, D., Dugat-Bony, E., et al. 2011, Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake, *FEMS Microbiol. Ecol.*, **77**, 533–45.
28. Reeve, J.N. 1992, Molecular biology of methanogens, *Annu. Rev. Microbiol.*, **46**, 165–91.
29. Klein, A., Allmansberger, R., Bokranz, M., Knaub, S., Müller, B. and Muth, E. 1988, Comparative analysis of genes encoding methyl coenzyme M reductase in methanogenic bacteria, *Mol. Gen. Genet.*, **213**, 409–20.
30. Dugat-Bony, E., Missaoui, M., Peyretaillade, E., et al. 2011, HiSpOD: probe design for functional DNA microarrays, *Bioinformatics*, **27**, 641–8.
31. Mihajlovski, A., Alric, M. and Brugere, J.F. 2008, A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the mcrA gene, *Res. Microbiol.*, **159**, 516–21.
32. Staden, R. 1996, The Staden sequence analysis package, *Mol. Biotechnol.*, **5**, 233–41.
33. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets, *Bioinformatics*, **27**, 863–4.
34. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
35. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. and Knight, R. 2011, UCHIME improves sensitivity and speed of chimera detection, *Bioinformatics*, **27**, 2194–200.
36. Larkin, M.A., Blackshields, G., Brown, N.P., et al. 2007, Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**, 2947–8.
37. Gouy, M., Guindon, S. and Gascuel, O. 2009, SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building, *Mol. Biol. Evol.*, **27**, 221–4.
38. Li, W. and Godzik, A. 2006, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658–9.
39. Luton, P.E., Wayne, J.M., Sharp, R.J. and Riley, P.W. 2002, The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill, *Microbiology*, **148**, 3521–30.
40. Studier, J. and Keppler, K. 1988, A note on the neighbor-joining algorithm of Saitou and Nei, *Mol. Biol. Evol.*, **5**, 729–31.
41. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406–25.
42. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.*, **28**, 2731–9.
43. Livak, K.J. and Schmittgen, T.D. 2001, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method, *Methods*, **25**, 402–8.
44. Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
45. Brauer, S.L., Cadillo-Quiroz, H., Yashiro, E., Yavitt, J.B. and Zinder, S.H. 2006, Isolation of a novel acidiphilic methanogen from an acidic peat bog, *Nature*, **442**, 192–4.
46. Dridi, B., Fardeau, M.L., Ollivier, B., Raoult, D. and Drancourt, M. 2012, *Methanomassiliicoccus luminyensis* gen. nov. sp. nov. a methanogenic archaeon isolated from human faeces, *Int. J. Syst. Evol. Microbiol.*, **62**, 1902–7.
47. Borrel, G., Harris, H.M.B., Tottey, W., et al. 2012, Genome sequence of '*Candidatus Methanomethylophilus alvus*' Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens, *J. Bacteriol.*, **194**, 6944–5.
48. Palmer, C. 2006, Rapid quantitative profiling of complex microbial populations, *Nucleic Acids Res.*, **34**, e5.
49. Fricke, W.F., Seedorf, H., Henne, A., et al. 2005, The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis, *J. Bacteriol.*, **188**, 642–58.
50. Borrel, G., Joblin, K., Guedon, A., et al. 2011, *Methanobacterium lacus* sp. nov., a novel hydrogenotrophic methanogen from the deep cold sediment of a meromictic lake, *Int. J. Syst. Evol. Microbiol.*, **62**, 1625–1629.
51. Summerer, D., Wu, H., Haase, B., et al. 2009, Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing, *Genome Res.*, **19**, 1616–21.
52. Lehours, A.C., Bardot, C., Thenot, A., Debroas, D. and Fonty, G. 2005, Anaerobic microbial communities in Lake Pavin, a unique meromictic lake in France, *Appl. Environ. Microbiol.*, **71**, 7389–400.
53. Jeon, S., Bunge, J., Leslin, C., Stoeck, T., Hong, S. and Epstein, S.S. 2008, Environmental rRNA inventories miss over half of protistan diversity, *BMC Microbiol.*, **8**, 222.
54. Militon, C., Rimour, S., Missaoui, M., et al. 2007, PhylArray: phylogenetic probe design algorithm for microarray, *Bioinformatics*, **23**, 2550–7.
55. Parisot, N., Denonfoux, J., Dugat-Bony, E., Peyret, P. and Peyretaillade, E. 2012, KASpOD – a web service for highly specific and explorative oligonucleotide design, *Bioinformatics*, **28**, 3161–3162.
56. McCarthy, A. 2010, Third generation DNA sequencing: pacific biosciences' single molecule real time technology, *Chem. Biol.*, **17**, 675–6.
57. Schadt, E.E., Turner, S. and Kasarskis, A. 2010, A window into third-generation sequencing, *Hum. Mol. Genet.*, **19**, R227–240.

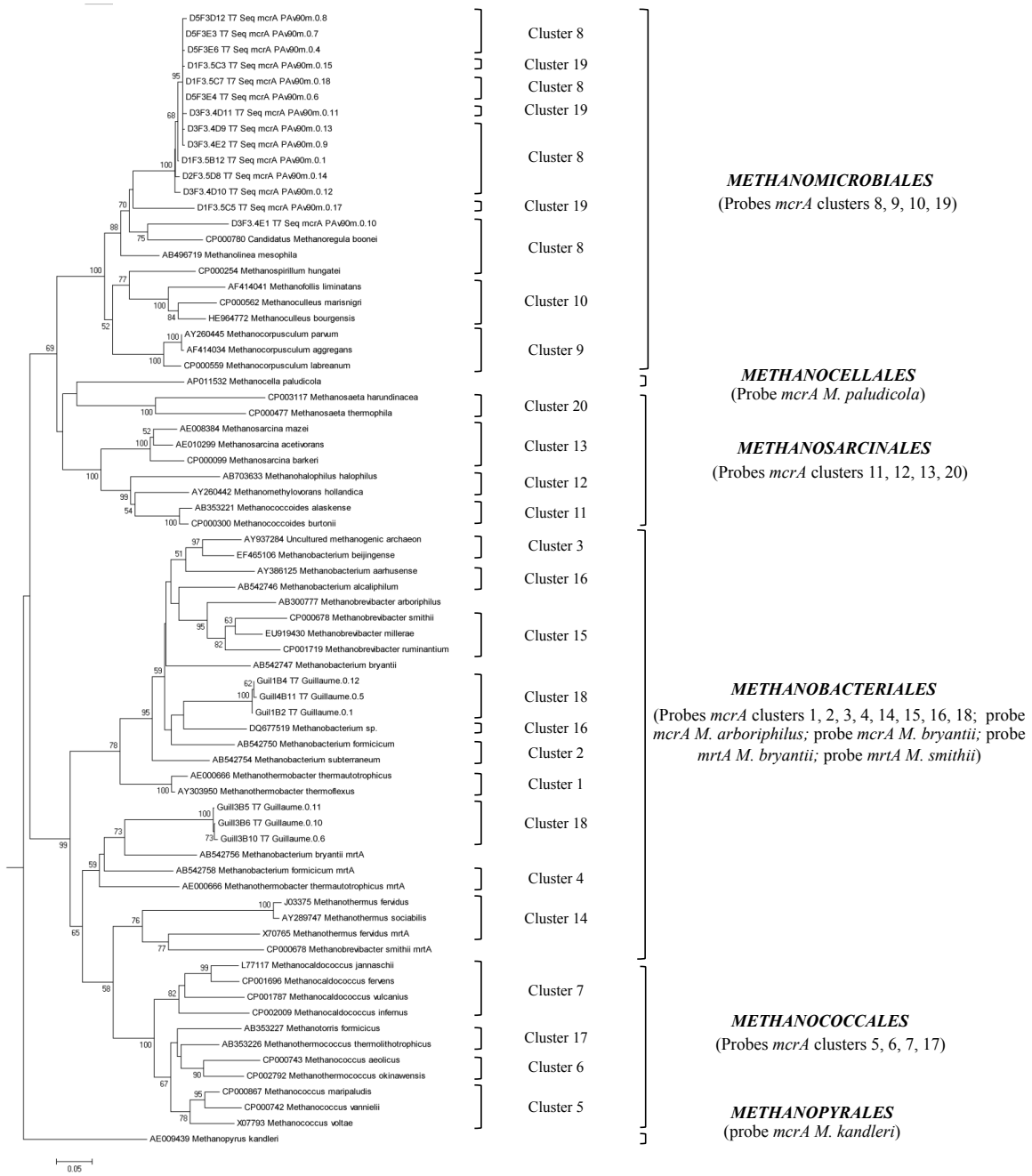


Fig. S1. Phylogenetic analysis of *mcrA* gene sequences used for the design of the second set of capture probes. The tree pointed out the probe coverage (indicated in brackets) within the orders *Methanomicrobiales*, *Methanocellales*, *Methanosarcinales*, *Methanobacteriales*, *Methanococcales* and *Methanopyrales*). Evolutionary history was inferred using the neighbour-joining method[38,39] (NJ, Poisson distance model) using Seaview software[35]. The final tree was drawn in MEGA 5[40]. The bars represent a 5% sequence divergence. Numbers at the nodes represent bootstrap values >50% (1,000 resamplings).

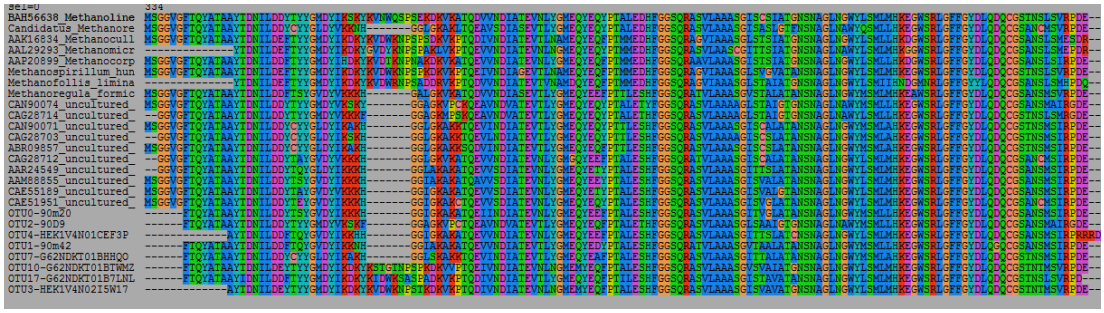


Fig. S2. Proteic alignment showing insertions events within the *mcrA* gene between cultured methanogen species and OTUs 10, 17 and 3 belonging to *Methanomicrobiales* order

	Primers / Probes (5' - 3')	Name	Size (bases)	Target	Reference
PCR	TAYATGTCNGGYGGTGTHGG	MM_01	20	*mcrA*	31
	ACRTTCATNGCRTAGTTNGG	MM_02			
	CCATCTCATCCCTGCGTGTC	454 Ti-A	20	Pyrosequencing adaptators	Roche Applied Science
	CCTATCCCCTGTGTGCCTTG	454 Ti-B			
	CCATCTCATCCCTGCGTGTCTCCGACGACTACACGACGACTTAYATGTCNGGYGGTGTHGG	Fusion primers	61	adapter A + RL001 MID + MM_01	This study
	CCTATCCCCTGTGTGCCTTGGCAGTCGACTACRTTCATNGCRTAGTTNGG		50	adapter B+ MM_02	
	CGATGTCCATCAGGCCCGA	50-68-Forwrad	19	*mcrA-fmd* spanning fragment	This study
	AGCTCGAAGTGGAAGGCACAA	97-117-Reverse	21		
Capture	TCTGGCTCGGATCCTACATGTCCGGTGGTGTCGGGTTCACCCAGTATGCA	P1	50	*mcrA*	This study
	CTGGTCTCTCCGGCTGGTACCTCTCCATGTATGTCCACAAGGAAGCATGG	P2			
	TGAAGACCACTTCGGTGGATCCCAGAGAGCAACCGTGCTCGCAGCTGCAT	P3			
	TCGGTCACCTCTCAGACATCGTCCAGACAAGCCGTGTATCCAAAGACCCC	P4			
	AAATTCCTGAGACTCGGCCCTGAACAGGATGCAAGAAAGCAGGAAATGAT	P5			
	CGATGATGCACATGGGTGCCCCTCTCGGTCAGCGTGCAATCACTCCTTAC	P6			
Pyrosequencing	ACACGACGACT	RL001 MID	11	-	Roche Applied Science
	ACACGTAGTAT	RL002 MID			
	ACACTACTCGT	RL003 MID			
	CCATCTCATCCCTGCGTGTCTCCGACGACT	A adapter-key	30	-	Roche Applied Science
	CCTATCCCCTGTGTGCCTTGGCAGTCGACT	B adapter-key			

Table S1. Primer and probes sets used for *mcrA* gene surveys

Probe name	Sequence (5'-3')
mcrA M. kandleri (1)	TCTACGACCAGATCTGGCTAGGATCCTACATGTCAGGAGGTGTCGGTTTC
mcrA M. paludicola (1)	TGTATGACCAGATCTGGCTCGGCTCCTACATGTCCGGTGGTGTCGGCTTC
mrtA M. smithii (1)	TATATGATCAGGTTTGGTTAGGTTCTTACATGTCAGGAGGTGTAGGTTTC
mrtA_M. bryantii (1)	TATACGATCAGATCTGGCTCGGATCTTACATGTCTGGTGGTGTTGGATTC
mcrA M_arboriphilus (1)	TATACGACCAAATTTGGTTAGGTTCTTACATGTCTGGTGGTGTTGGATTT
mcrA_M. bryantii (1)	TTTACGACCAAATCTGGCTTGGTTCATACATGTCAGGTGGTGTAGGATTC
mcrA cluster 1 (1)	CAGTGTGGTGCATCCAACGTCTTCTCAATAAGGGGCGACGAGGGACTGCC
mcrA cluster 2 (2)	ACTGGAAATGATGAAATCGCTGATGAAAT**Y**GACCAGAGATACGTCCTTAA
mcrA cluster 3 (2)	GCTGCAGCATCTGCATGTTCCACTGGATTTGCAACTGGAAACGC**M**CAAAC
mcrA cluster 4 (4)	GCAGGTGAAGCAGCAAT**Y**GCTGACTTCTCATACGC**W**GCAAAACACGCCGA
mcrA cluster 5 (4)	GGTAGAGTATGTGACGG**Y**GGTACAAT**Y**TCAAGATGGTCTGCAATGCAGAT
mcrA cluster 6 (4)	TCAGTATGTATGGCAACAGGAAACTCAAATGC**W**GG**R**GTTAATGGATGGTA
mcrA cluster 7 (4)	ACAATAGCAAGATGGAGTGC**W**ATGCAGAT**W**GGAATGTCATTCATTACAGC
mcrA cluster 8 (8)	TGCACAAGGAAGG**M**TGGTCACGTCTCGG**M**TTCTTCGG**M**TACGACCTGCAG
mcrA cluster 9 (1)	CAGTATGAACAGTTCCCGACCATGATGGAAGACCACTTCGGCGGTTCCCA
mcrA cluster 10 (4)	CAGTACGAGCAGTTCCCGAC**S**ATGATGGA**R**GACCACTTCGGCGGGTCCCA
mcrA cluster 11 (1)	ATGGCTGATATCATTCAGACAAGCCGTGTTGACGCAGAAGATCCAGCACA
mcrA cluster 12 (2)	CCCTTGAGGTAGTCGGTGCAGG**M**TGTATGCTCTACGACCAGATCTGGCT
mcrA cluster 13 (1)	GTTCTGTCCTACCAGGGCGACGAAGGTCTCCCAGACGAACTCCGTGGTCC
mcrA cluster 14 (8)	TAGCAACCGAAGTTACACTTTA**YR**GTCTTGA**M**CAATATGAAGAATATCCA
mcrA cluster 15 (2)	CATTAGAACAATACGAAGAATACCCAGCTTTACT**Y**GAAACTCACTTCGGT
mcrA cluster 16 (8)	TGTGATGGTGGTAC**M**AC**W**TCCCGATGGTCTGCTATGCAGAT**Y**GGTATGTC
mcrA cluster 17 (2)	GCAATGCAGATAGGGATGTCATTCATTACAG**Y**ATACAAACTCTGTGCTGG
mcrA cluster 18 (8)	TATACGATCAGATCTGGCTAGGTTC**W**TACATGTCAGGTGG**W**GTAGG**W**TTC
mcrA cluster 19 (4)	CGGTGGTGTCGGTTTCACCCAGTATGCAAC**M**GC**W**GCATACACCGACAACA
mcrA cluster 20 (2)	ATCCGAACTACGC**S**ATGAACGTCGGCCACCAGGGCGAGTATGCAGGCATC

Table S2. Oligonucleotide probes sequence targeting the Methyl Coenzyme M reductase subunit A gene (*mcrA*).

The 49 and 50-mers probes designed could be specific (1 oligonucleotide) or degenerated (2, 4, 8 oligonucleotides) as indicated in brackets. Probes were designed from the most conserved regions of each group determined after a clustering using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/)

	Primers (5' - 3')	Name	Size	Target	Reference
qPCR (Enrichment calculation)	TGCAAGGGCACATGCAACAC	346-365-Forward	20	*mcrA*	This study
	TGCTGCAAATCTGGGCACTG	516-535-Reverse			
	GCTGCTTATGTGGCCTGGAT	55-174-Forward	20	*fmdA*	This study
	GCATACCGAGGCGTTCGTT	326-344-Reverse	19		
qPCR (Methanogen abundance)	CAGGCTGTTCAACCGCATTTGC	1F45_1_212-233	22	OTU78	This study
	TCAGACCTTCATCGCTTCTGAT	1R30_1_364-385	22		
	GCTTCCCGGCCGCAATGGA	1F67_1_181-199	19	OTU13	This study
	TTGACACCAGCGTTCGCGT	1R57_1_277-295	19		
	CCCAGAGAGCATCCGTTCTG	1F101_1_229-248	20	OTU9	This study
	CAAGCGTCCCCAGCCTTCC	1R29_1_338-356	19		
	TGCAACTGAAGTCACGCTCTACG	1F8_1_136-158	23	OTU2	This study
	GGACAGACCTGATGCGGCT	1R54_1_235-253	19		
	CGAGAGCCACTTCGGCGGA	1F68_1_211-229	19	OTU7	This study
	TTCCTTGTGGGCGAGCATGG	1R59_1_324-343	20		
	CTTCGGTGGTTCCCAGCGTGCAT	1F93_1_224-246	23	All	This study
	TGCAGGTCGTAGCCGAAGAAGC	1R24_1_364-385	22		

Table S3. Primers sets used for qPCR experiments

2.4 Application à d'autres biomarqueurs

L'approche de capture de gènes en solution a récemment été appliquée à d'autres gènes biomarqueurs pour l'étude de différents environnements. Ainsi, les sondes de capture *mcrA* sont actuellement utilisées pour explorer la diversité des archées méthanogènes chez le ruminant. Ce projet, nommé CREDIT, est financé par l'Agence Nationale de la Recherche (ANR) et porté par le Dr. Diego Morgavi (INRA Clermont-Ferrand Theix). Il vise au développement de nouvelles stratégies alternatives afin de limiter la production de méthane chez le ruminant.

De même, l'adaptation des communautés microbiennes eucaryotes à différentes contraintes environnementales a pu être étudiée en collaboration avec l'équipe du Dr. Roland Marmeisse (UMR CNRS 5557 Ecologie Microbienne). Au travers de l'Initiative Structurante EC2CO (« Ecosphère Continentale et Côtière ») portée par le Dr. Patricia Luis, une approche de capture de gènes a été mise en place sur différents biomarqueurs fonctionnels et phylogénétiques : des protéines riches en cystéine (CRP), des glycoside hydrolases (GH), des peroxydases (DYP), et enfin le gène codant l'ARNr 18S. La capture des GH a été effectuée à partir d'ADN complémentaires et a permis l'obtention de séquences fonctionnelles, non encore décrites dans les bases de données, qui ont pu être exprimées avec succès chez *Saccharomyces cerevisiae* (ANNEXE 1). La détermination des sondes oligonucléotidiques de capture a été effectuée grâce au logiciel KASpOD (Parisot *et al.* 2012).

Une autre application intéressante de la capture de gènes en écologie microbienne concerne le biomarqueur le plus étudié : le gène ADNr 16S. C'est avec cet objectif qu'un premier jeu de sondes généralistes a été déterminé avec KASpOD et est actuellement testé sur divers environnements : symbiome microbien d'arthropodes (collaboration avec le Dr. Sylvain Charlat de l'UMR CNRS 5558 LBBE et le Dr. Jean-Christophe Simon de l'INRA de Rennes), rumen (collaboration avec le Dr. Diego Morgavi de l'INRA de Clermont-Ferrand Theix), stations d'épuration (collaboration avec le Dr. Denis Le Paslier du Genoscope), et environnements lacustres (collaboration avec le Dr. Corinne Biderre-Petit de l'UMR CNRS 6023 LMGE). Les premiers résultats de séquençage montrent des taux d'enrichissements en séquences d'ADNr 16S de plus de 90% sur certains échantillons, avec la possibilité de reconstruire des fragments génomiques de plus de 2 kpb permettant ainsi l'étude du biomarqueur dans son intégralité mais donnant également accès au gène ADNr 23S. La diversité taxonomique observée semble également moins biaisée que celle retrouvée par les

approches de séquençage haut-débit d'amplicons. De tels résultats soulignent donc la pertinence de l'approche de capture de gènes en écologie microbienne pour permettre l'étude des communautés microbiennes. En effet, la longueur des fragments obtenus permet une affiliation taxonomique plus précise, l'absence de PCR permet de s'affranchir des biais inhérents à cette technique, et l'enrichissement est tel qu'il permet l'étude de la biosphère rare.

En parallèle de l'enrichissement par capture de gènes, une perspective intéressante repose sur les approches de soustraction. Les études transcriptomiques et métatranscriptomiques par séquençage massif sont de plus en plus employées en écologie microbienne. Néanmoins, les données de séquençage sont généralement polluées par un pourcentage significatif de séquences d'ADNr et ce malgré la multiplication des kits commerciaux de déplétion. C'est pourquoi, un jeu de sondes généralistes ciblant les gènes ADNr 16S, 18S, 23S et 28S a été déterminé en utilisant KASpOD et appliqué pour la déplétion d'échantillons environnementaux en séquences ribosomiques. En collaboration avec l'INRA de Clermont-Ferrand Theix, le Dr. Alain Sarniguet de l'INRA de Rennes et le Dr. Patrick Mavingui de l'UMR CNRS 5557 Ecologie Microbienne, ces approches soustractives sont en cours de validation sur des échantillons variés comme le rumen, le complexe ectomycorhizien ou le microbiote du moustique tigre *Aedes albopictus*.

2.5 Discussion

L'approche moléculaire de capture de gènes a démontré sa pertinence pour assurer un enrichissement significatif des séquences ciblées. En effet, lors des différentes études nous avons pu obtenir une efficacité d'enrichissement supérieure à 90% du biomarqueur ciblé en partant d'un échantillon métagénomique complexe. L'étude comparative portant sur le gène *mcrA* a clairement démontré les difficultés à décrire la diversité de façon exhaustive par le séquençage direct de l'ADN métagénomique. Afin d'avoir une vision globale de la diversité des communautés méthanogènes au sein de cet environnement, des millions de lectures supplémentaires auraient été nécessaires, en accord avec les conclusions tirées des travaux initiés par Quince *et al.* (2008) sur les efforts de séquençage à fournir pour explorer la diversité microbienne.

Cette stratégie a aussi permis, contrairement aux approches classiques utilisant la PCR, d'explorer de manière plus exhaustive les communautés microbiennes. Un tel résultat a été rendu possible grâce au fait que cette approche s'affranchit des biais occasionnés par les

méthodes basées sur la PCR. En effet, l'efficacité de ces dernières est intimement liée au choix des amorces et à la part relative de chaque communauté à identifier au sein de l'écosystème. Ainsi, l'utilisation de la capture a permis d'identifier de nouvelles séquences non répertoriées dans les bases de données mais également d'accéder à des populations rares. Une autre limitation des approches PCR, levée par l'approche capture, est la taille des séquences pouvant être identifiées. En effet, la longueur des amplicons obtenus n'est pas toujours suffisante pour caractériser précisément les communautés microbiennes (Wommack *et al.* 2008). L'approche capture, permettant quant à elle d'obtenir des fragments de grande taille, un nombre de sites moléculaires du biomarqueur ciblé plus important est donc disponible pour entreprendre l'identification phylogénétique des communautés.

La possibilité d'identifier de grandes régions d'ADNg représente donc l'autre atout majeur de cette approche de capture. L'émergence du séquençage de troisième génération devrait d'ailleurs permettre de faciliter l'obtention de données sur de très grandes régions d'ADN capturées et améliorer encore nos connaissances sur le monde microbien. Capturer un fragment génomique contenant les deux types de biomarqueurs, phylogénétique et fonctionnel, permettrait ainsi de relier structure et fonction des écosytèmes microbiens et répondre à la question principale en écologie microbienne : « qui fait quoi ? ». De plus, en ce qui concerne les biomarqueurs fonctionnels, les régions capturées peuvent inclure plusieurs gènes pouvant faire partie de la même unité transcriptionnelle et donc permettre de donner des pistes pour l'annotation fonctionnelle de nouveaux gènes à séquences inconnues mais associés à des gènes codant pour des protéines à fonction connue (Overbeek *et al.* 1999 ; Korbel *et al.* 2004). Une telle approche représente donc une alternative pour la prédiction de la fonction des gènes. Actuellement, la prédiction de fonction des gènes est généralement basée sur une recherche d'homologie de séquences dans les bases de données en utilisant notamment les outils BLAST (Altschul *et al.* 1990) mais en gardant à l'esprit qu'un grand nombre de séquences sont mal annotées (Schnoes *et al.* 2009). De même, du fait de la connaissance partielle de l'extraordinaire diversité des microorganismes dans les environnements, il a été montré que suite au séquençage direct de métagénomes, environ 30 à 60% des protéines ne pouvaient être clairement identifiées avec une fonction connue en utilisant les bases de données actuelles (Vieites *et al.* 2009).

3. Développement d'un outil d'affiliation taxonomique et fonctionnelle des séquences métagénomiques : AFFILGOOD

3.1 Contexte

La métagénomique, en permettant l'étude directe du matériel génétique d'écosystèmes complexes, constitue l'une des approches les plus efficaces pour pouvoir appréhender la structure et le fonctionnement des communautés microbiennes. Néanmoins, son couplage aux méthodes de séquençage haut-débit pose de nouvelles questions techniques et méthodologiques, notamment pour permettre l'affiliation taxonomique des masses de données générées.

La procédure d'affiliation taxonomique la plus complète repose sur la recherche, au sein des bases de données, des séquences similaires à la séquence requête (*e.g.* BLAST). Il s'agit ensuite de sélectionner un ensemble de candidats en fonction de la qualité d'alignement avec la séquence requête pour construire un alignement multiple. Finalement, en appliquant une méthode de reconstruction phylogénétique, cet alignement permet l'élaboration d'un arbre assurant l'affiliation taxonomique de la séquence étudiée. Néanmoins, une part importante des séquences présentes dans les bases de données internationales provient d'espèces cultivables. Or, dans un métagénome, issu d'un environnement complexe, certaines séquences correspondent à de nouvelles espèces encore non identifiées dont certaines peuvent ne présenter que peu de similarités avec les séquences des bases de données. Pour prendre en compte la totalité des séquences de marqueurs phylogénétiques générées par les projets de métagénomique, une autre grande classe de méthodes a donc vu le jour. Il s'agit des méthodes qui vont évaluer la composition des séquences pour en assurer la classification. En associant l'étude de la fréquence d'oligomères courts (*i.e.* moins de 10 nucléotides) avec des modèles statistiques, il a été possible de concevoir des algorithmes d'affiliation taxonomique tels que RDP Classifier (Wang *et al.* 2007). Cependant la taille des séquences est un paramètre critique pour cette famille de méthodes puisqu'aucune d'entre elles ne fonctionne réellement efficacement sur des séquences de moins de 1 kpb en raison du nombre limité d'oligomères qu'elles contiennent (McHardy *et al.* 2007). Par ailleurs, cette famille de méthodes est très sensible aux erreurs de séquençage ainsi qu'aux transferts horizontaux de gènes (McHardy *et al.* 2007 ; Dröge & McHardy 2012).

Le défi actuel consiste donc à mettre au point des méthodes plus probantes garantissant une classification pertinente des séquences courtes issues des NGS.

3.2 Objectif

L'objectif de ce travail de recherche vise donc à développer un outil d'affiliation taxonomique et fonctionnelle des données produites par les techniques de séquençage haut-débit. L'approche envisagée tire parti de la précision des méthodes basées sur la similarité de séquences mais en se focalisant uniquement sur des signatures oligonucléotidiques courtes et hautement spécifiques afin de réduire la complexité du traitement. En effet, l'ensemble des signatures disponibles est comparé au jeu de données de séquences à affilier. Toute séquence possédant une ou plusieurs signatures d'un même taxon ou gène fonctionnel sera affiliée à celui-ci. Dans le cas où plusieurs signatures de groupes taxonomiques ou fonctionnels différents sont identifiées pour une même lecture, le conflit est résolu par la reconstruction d'un arbre phylogénétique avec les groupes impliqués. L'ensemble de la démarche mise en place a donné naissance au logiciel nommé AFFILGOOD.

3.3 Principaux résultats

Dans un premier temps, les développements se sont portés sur l'affiliation taxonomique des séquences. Comme pour de nombreuses méthodes d'affiliation basées sur le gène ADNr 16S, la première étape consiste à disposer d'une base de données de séquences de haute qualité en terme de séquences mais également en terme d'affiliation taxonomique.

3.3.1 Construction de la base de données de séquences d'ADNr 16S

A partir des divisions procaryotes (PRO) et environnementales (ENV) de la base de données EMBL, les séquences correspondant au gène ADNr 16S ont été filtrées grâce à l'utilisation de mots-clés. Seules les séquences ayant une taille comprise entre 1200 et 1600 nt sont ensuite sélectionnées pour ne conserver que des séquences complètes. La qualité des séquences est également un critère important pour pouvoir disposer d'un outil d'affiliation taxonomique fiable. Ainsi, les séquences contenant plus de 1% de bases indéterminées (*e.g.* N) ou plus de 5 bases indéterminées consécutives sont exclues. A la fin de cette étape, il est possible d'organiser ces données de séquences d'un point de vue taxonomique en utilisant la taxonomie du NCBI et les informations contenues dans le champ OC (*Organism Classification*) des fiches EMBL récupérées. Pour chaque genre recensé dans la base de données, l'ensemble des séquences qui lui sont affiliées est *clusterisé* de manière stricte (*i.e.*

100% d'identité) pour éliminer toute redondance d'information. Enfin, il est nécessaire de s'assurer de l'affiliation taxonomique fournie dans les fiches EMBL. Pour cela, nous avons choisi d'utiliser la base de données de séquences d'ADNr SILVA (Quast *et al.* 2013) à partir de laquelle une séquence de référence par genre a été sélectionnée. L'évaluation de la qualité de l'affiliation des séquences s'effectue alors par comparaison de séquences deux à deux en utilisant une version modifiée de l'algorithme ClustalW2 (Larkin *et al.* 2007) (**Figure 12**). Au final, l'approche mise en place a permis de constituer une base de données de 66 075 séquences d'ADNr 16S correspondant à 2069 genres procaryotes différents. Par ailleurs, il s'agit de la base de données utilisée par l'algorithme de détermination de sondes oligonucléotidiques PhylGrid 2.0 (Jaziri *et al.* 2014b).

Afin de limiter la complexité de certaines étapes de l'algorithme AFFILGOOD, une seconde version de la base de données a été construite pour réduire le nombre de séquences par genre. Ainsi, pour chaque genre étudié, seules 10 séquences représentatives de la diversité sont conservées. Pour les genres avec plus de 10 séquences, une matrice de similarités entre toutes les séquences est donc établie grâce au logiciel ClustalW2 modifié pour ne réaliser que l'alignement des paires de séquences. A partir de cette matrice, et pour chaque séquence, la moyenne de toutes ses valeurs de similarités avec les autres séquences est calculée. La séquence avec la moyenne la plus élevée (M_1) est conservée car il s'agit de la séquence la plus proche de toutes les autres. De même, la séquence la plus éloignée des autres est récupérée (avec la valeur de moyenne la plus faible (M_2)). Puis, l'écart entre ces deux valeurs extrêmes de moyenne est divisé par 8 afin de déterminer un pas (noté d) pour la sélection des 8 séquences restantes : $d=(M_1-M_2)/8$. Ainsi, la séquence avec la moyenne la plus proche de M_2+d est conservée, puis celle la plus proche de M_2+2d et ce jusqu'à M_2+8d. Finalement, tout en conservant la diversité de séquences au sein de chaque genre procaryote (**Figure 13**), cette version de la base de données ne contient que 10 781 séquences.

3.3.2 Détermination des signatures taxonomiques

En plus de leur grande spécificité envers les taxons étudiés, la particularité des signatures utilisées repose sur leur pouvoir exploratoire. En effet, il est nécessaire qu'elles puissent permettre d'identifier les communautés bactériennes pour lesquelles aucune séquence n'est encore disponible. Cependant, contrairement aux sondes définies pour les expérimentations de biopuces ADN, leur utilisation *in silico* ne nécessite pas de prendre en compte les critères thermodynamiques, la dégénérescence ou leur taille lors de leur

détermination. En revanche, afin de conserver un fort pouvoir discriminant, leur spécificité apparaît comme le critère majeur à considérer. Le logiciel KASpOD (Parisot *et al.* 2012) a donc été choisi pour déterminer un premier ensemble de signatures à différents niveaux taxonomiques allant du genre au phylum. Ainsi, pour chaque taxon étudié différentes tailles de sondes (*i.e.* de 17 à 31-mers) ont été testées afin de déterminer les signatures les plus spécifiques.

Ces résultats sont toujours en cours d'analyse pour pouvoir permettre d'établir le jeu de signatures le plus efficace pour l'affiliation taxonomique des séquences métagénomiques.

3.3.3 Affiliation taxonomique

Afin de procéder à la classification taxonomiques des séquences métagénomiques possédant le biomarqueur étudié, l'ensemble des signatures disponibles est comparé au jeu de données de séquences à affilier en utilisant l'outil PatMaN (Prüfer *et al.* 2008). Cet algorithme permet la recherche rapide et sensible de similarités entre des sondes dégénérées et d'importantes masses de données de séquences.

A partir de ces résultats, il est possible de déterminer pour chaque séquence à affilier la liste des signatures qu'elle possède. Si la séquence ne présente que des signatures d'un même taxon elle sera alors affiliée à celui-ci. En revanche, dans le cas où une lecture s'aligne significativement avec plusieurs signatures de taxons différents, un arbre phylogénétique est construit. Pour cela, un alignement multiple est réalisé entre les séquences d'ADNr 16S des différents taxons incriminés et la séquence requête grâce à l'utilisation du logiciel Muscle (Edgar 2004). Afin de réduire les temps de calculs, c'est la base de données réduite qui est utilisée pour cette étape. Puis, par l'intermédiaire du logiciel trimAl (Capella-Gutierrez *et al.* 2009), tous les sites moléculaires avec *gaps* sont supprimés, à moins que cela n'en supprime plus de 80%. Dans ce cas, les sites moléculaires avec le moins de *gaps*, et donc les plus informatifs, sont considérés pour conserver au moins 20% des sites de l'alignement multiple initial.

A partir de cet alignement multiple, un arbre phylogénétique est reconstruit *via* le logiciel FastTree (Price *et al.* 2010). FastTree repose sur le principe d'évolution minimale (*i.e.* méthode de distance qui choisit l'arbre le plus parcimonieux parmi l'ensemble des arbres possibles) en proposant une heuristique proche de l'algorithme de *Neighbor-Joining* (Saitou & Nei 1987) pour construire l'arbre. La topologie de l'arbre, et notamment les distances

phylogénétiques entre les différentes feuilles est ensuite optimisée par l'application d'une méthode du maximum de vraisemblance. Le modèle d'évolution nucléique utilisé est celui de Jukes et Cantor (Jukes & Cantor 1969).

Cet arbre est ensuite parcouru afin de déterminer la position de notre séquence requête. La composition taxonomique du nœud contenant la séquence requête permettra d'assurer au mieux l'affiliation de cette séquence.

3.4 Discussion

L'algorithme AFFILGOOD représente donc une approche innovante pour l'affiliation taxonomique des séquences métagénomiques basée sur l'utilisation de signatures génomiques établies à différents niveaux taxonomiques à partir du gène ADNr 16S. La détermination de ces signatures, pouvant être exploratoires, s'effectue par l'intermédiaire du logiciel KASpOD et l'exploitation des séquences d'une base de données propriétaire dont la construction a été optimisée. Les résultats préliminaires s'avèrent concluant quant à la qualité des affiliations proposées en comparaison avec deux autres logiciels faisant référence dans ce domaine (*i.e.* RDP (Wang *et al.* 2007) et STAP (Wu *et al.* 2008)).

Cette approche, toujours en cours de développement, peut faire l'objet de plusieurs améliorations. C'est par exemple le cas lorsque plusieurs signatures de groupes taxonomiques ou fonctionnels différents sont détectées dans une même séquence. Afin de remplacer l'étape de reconstruction phylogénétique *de novo* qui peut s'avérer chronophage, il est envisagé d'intégrer le placement de la séquence métagénomique dans un arbre phylogénétique préétabli, en utilisant un outil comme pplacer (Matsen *et al.* 2010), afin de diminuer les temps de calculs. Le logiciel pplacer permet de placer un ensemble de séquences, mêmes partielles, dans un arbre phylogénétique préexistant et propose un calcul de vraisemblance permettant d'estimer la qualité de placement dans cet arbre. Basé sur l'utilisation conjointe d'une méthode de maximum de vraisemblance et d'une méthode bayésienne, cet outil fournira assurément de meilleurs résultats que l'approche actuellement implémentée.

Les stratégies haut-débit basées sur l'utilisation de *k*-mers sont de plus en plus employées pour l'étude des données métagénomiques. Outre les assembleurs qui utilisent conjointement graphes et *k*-mers, il existe déjà des stratégies de classification et d'annotation, taxonomique ou fonctionnelle, des séquences métagénomiques (Zhu *et al.* 2010 ; Nalbantoglu *et al.* 2011 ; Niu *et al.* 2011 ; Edwards *et al.* 2012 ; Jiang *et al.* 2012 ; Wang *et al.* 2014b ; Tu

et al. 2014a). Tu et collaborateurs (Tu *et al.* 2014a) proposent ainsi une approche semblable à AFFILGOOD pour assurer l'identification taxonomique de souches, dont les génomes sont entièrement, séquençés, au sein de métagénomes à partir de signatures génomiques.

Si toutefois la restitution taxonomique obtenue par l'étude du gène codant pour l'ARNr 16S n'était pas suffisante, il est également envisagé d'orienter notre stratégie d'étude de la biodiversité microbienne vers l'analyse d'autres marqueurs phylogénétiques tels que les gènes de ménage (*recA*, *rpoB*, etc. (Liu *et al.* 2012b)). De même, cette stratégie pourra être transposée à l'étude des microorganismes eucaryotes *via* la détermination de signatures basées sur le gène codant pour l'ARNr 18S.

La classification fonctionnelle des données de métagénomique étant tout aussi importante (Steele *et al.* 2009 ; Mitra *et al.* 2011 ; Morales & Holben 2011 ; Prakash & Taylor 2012), un intérêt particulier sera porté à la détermination de signatures fonctionnelles. L'outil AFFILGOOD devra alors être validé pour ce volet fonctionnel en utilisant un jeu de signatures exhaustif. Ainsi, une détermination de signatures pourra être effectuée pour de nombreux types de gènes fonctionnels, qu'il s'agisse de gènes de ménages pour améliorer la résolution taxonomique, de gènes de virulence pour la détection de microorganismes pathogènes ou des gènes biomarqueurs de voies métaboliques d'intérêt comme la dégradation de polluants.

Conclusion et perspectives

De par leur rôle dans les grands cycles biogéochimiques, les microorganismes sont des acteurs extrêmement importants dans l'équilibre et le fonctionnement des écosystèmes. Leur diversité et leur capacité évolutive, conséquences entre autres des échanges d'informations génétiques, confèrent à ces microorganismes des potentialités adaptatives insoupçonnées. Ainsi, les bactéries peuvent occuper de nombreuses niches écologiques dont les plus extrêmes. Certaines possèdent également des caractéristiques enzymatiques d'intérêt comme la synthèse de molécules à haute valeur ajoutée ou la dégradation de polluants. Néanmoins, l'exploitation de ces potentialités métaboliques nécessite une meilleure connaissance du monde microbien.

A ce jour, la masse de données existantes concerne principalement des organismes modèles et nous sommes encore bien loin d'avoir une image exhaustive de la diversité biologique. L'écologie microbienne s'attache donc à accroitre nos connaissances du vivant en étudiant la biodiversité microbienne des écosystèmes au travers de l'utilisation de techniques adaptées. Au cours de ce travail de thèse, les avantages et les inconvénients de chacune des techniques actuellement disponibles en écologie microbienne ont pu être décrits. Il est apparu important de noter l'émergence de la génomique environnementale grâce à la révolution technologique apportée par le développement de nouvelles générations de séquençage. La génomique environnementale a modifié profondément et durablement les stratégies expérimentales mises en œuvre pour l'évaluation de la biodiversité taxonomique et fonctionnelle des écosystèmes. Cependant, le flux massif de données générées engendre de nouveaux verrous techniques et méthodologiques qu'il est parfois difficile de lever. En effet, l'augmentation constante du volume de production se fait au détriment de la qualité des données et des capacités d'analyse. Il est alors nécessaire de proposer de nouvelles stratégies d'exploration de la structure et de la fonction des communautés microbiennes.

Parmi les méthodes de réduction de complexité, on peut citer notamment le séquençage à haut-débit d'amplicons, les biopuces ADN ou les techniques de cellule isolée. En tirant parti du potentiel même du séquençage massif et des biopuces ADN, des stratégies de capture de gènes ont également vu le jour afin d'analyser finement et de manière ciblée certaines populations ou métabolismes d'intérêt. Il faut aussi noter qu'un grand nombre de ces approches de biologie moléculaire nécessite l'utilisation d'oligonucléotides comme sondes ou comme amorces. Il apparaît alors essentiel de disposer de logiciels performants pour une

détermination efficace de ces séquences oligonucléotidiques tel que nous l'avons présenté au travers de la synthèse bibliographique.

Toujours dans le but d'améliorer la détermination de ces séquences, nous avons été amenés à développer plusieurs logiciels de sélection de sondes oligonucléotidiques adaptés aux problématiques environnementales. Dans un premier temps, le logiciel PhylArray (Militon *et al.* 2007), premier logiciel de détermination de sondes exploratoires, a été amélioré pour le rendre applicable aux *designs* à grande échelle. Ainsi, l'algorithme PhylGrid 2.0 (Jaziri *et al.* 2014b) s'est appuyé sur la grille de calculs européenne pour assurer la détermination de 19 874 sondes ciblant le gène ADNr 16S de 2069 genres procaryotes. Malgré son caractère hautement parallélisable, PhylGrid 2.0 est rapidement limité par les jeux de données volumineux en raison de son étape d'alignement multiple des séquences. Afin de pallier cela, une nouvelle stratégie basée sur l'utilisation de *k*-mers a été mise au point. Ce logiciel, nommé KASpOD (Parisot *et al.* 2012), combine les critères de sensibilité, de spécificité et le caractère exploratoire pour la détermination d'oligonucléotides de qualité, et cette détermination peut être réalisée à partir de grands jeux de données. Ainsi, à partir d'un nombre trois fois plus important de séquences que PhylGrid 2.0, 56 613 sondes ciblant 1295 genres procaryotes ont pu être déterminées grâce à cette approche. Ces deux ensembles de sondes ont été rendu disponibles à la communauté scientifique *via* l'implémentation d'une base de données, PhylOPDb (Jaziri *et al.* 2014a). Ces séquences oligonucléotidiques sensibles, spécifiques et exploratoires, facilement accessibles, peuvent être utilisées pour diverses applications (*e.g.* PCR, FISH, biopuces ADN, capture de gènes). L'étude de la diversité fonctionnelle au sein des environnements n'a pas été délaissée avec le développement, toujours en cours, d'un troisième outil de détermination de sondes dédié aux gènes codant pour des protéines.

Tirant parti de cette expertise sur la détermination de sondes, les travaux menés au cours de cette thèse ont conduit au développement de nouvelles méthodes moléculaires et bioinformatiques pour l'exploration à haut-débit de la diversité taxonomique et fonctionnelle d'environnements complexes. Ainsi, deux biopuces environnementales ont été mises au point pour la caractérisation phylogénétique des communautés microbiennes. La première, nommée HuGChip (Tottey *et al.* 2013), est dédiée à l'étude du microbiote intestinal humain alors que la seconde, toujours en cours de validation, est une biopuce phylogénétique généraliste. Ces biopuces permettent l'analyse simultanée de plusieurs (*i.e.* jusqu'à 16) échantillons métagénomiques en proposant l'identification rapide de taxa déjà caractérisés ou non. Ces

travaux de thèse nous ont aussi amené à développer une nouvelle méthode d'étude des écosystèmes microbiens. Cette approche, appelée capture de gènes en solution, présente de nombreux avantages pour étudier la diversité des communautés microbiennes à partir d'échantillons environnementaux complexes. La capture offre la possibilité de cibler spécifiquement des populations microbiennes difficilement accessibles par d'autres techniques. En comparaison à une approche métagénomique directe ou à une approche amplicons, la capture a en effet montré son efficacité pour assurer une meilleure évaluation de la diversité microbienne, notamment des populations rares. Outre la possibilité de pouvoir explorer rapidement et de manière plus précise la diversité des communautés microbiennes à un très haut-débit, l'utilisation de la capture de gènes pour identifier de larges régions d'ADN génomique peut favoriser l'identification de nouveaux gènes mais aussi la compréhension de processus adaptatifs liés à l'environnement. Enfin, une nouvelle approche bioinformatique d'analyse taxonomique et fonctionnelle des données métagénomiques a été développée en tirant profit des sondes oligonucléotidiques précédemment développées. En effet, certaines sondes peuvent être qualifiées de signatures génomiques pour permettre l'affiliation taxonomique et/ou fonctionnelle des séquences. Cet outil, nommé AFFILGOOD, pourra permettre l'analyse haut-débit des données générées par les nouvelles techniques de séquençage.

Les résultats obtenus au cours de cette thèse ouvrent de nombreuses perspectives tant en bioinformatique qu'en écologie microbienne. La détermination de sondes reste au cœur de nombreuses techniques (*e.g.* FISH, PCR, biopuces ADN, criblage de banques métagénomiques et capture de gènes) et nécessite des logiciels pouvant gérer les flux massifs de données issues du séquençage de nouvelle génération. Chaque outil de détermination de sondes développé n'est donc pas guidé par une question biologique particulière mais permet de répondre à différentes problématiques de manière rapide et précise. Ainsi, ils peuvent s'appliquer à divers environnements allant du sol au microbiote intestinal humain en passant par les environnements lacustres. Le diagnostic rapide d'un environnement, par l'étude de sa composition taxonomique ou les voies métaboliques en présence est alors envisageable. Ces résultats permettent ainsi d'orienter, par exemple, des stratégies de bioremédiation d'environnements pollués, de mener des études en santé humaine, ou de détecter la présence de pathogènes (eucaryotes, procaryotes ou virus) par leur caractérisation directe ou l'identification de leurs gènes de virulence. Parmi les logiciels de détermination de sondes développés, KASpOD apparaît comme le plus représentatif de cette polyvalence. Bien

qu'appliqué au développement d'un jeu de sondes ciblant le gène ADNr 16S, cet algorithme est capable de travailler sur tout type de gène. Ainsi, il est aujourd'hui indissociable de l'approche de capture de gènes afin d'assurer l'exploration taxonomique et/ou fonctionnelle des communautés aussi bien procaryotes qu'eucaryotes au sein de tous les types d'environnements.

A l'instar de PhylGrid 2.0, KASpOD a apporté de nombreuses optimisations en terme de temps de calcul et pourrait évoluer vers le déploiement sur des architectures plus puissantes de type grilles de calcul. Actuellement, KASpOD est déployé sur un *cluster* composé de 140 CPUs hébergé au Centre Régional des Ressources Informatiques (CRRI) de Clermont-Ferrand. Une perspective intéressante serait de pouvoir bénéficier du potentiel des architectures hautement parallèles comme par exemple la grille de calcul européenne. Cette augmentation des capacités de calculs offre la possibilité de considérer un plus grand nombre de critères assurant la sélection de sondes performantes. Parmi ces critères, il est possible de citer les paramètres thermodynamiques ou les structures secondaires des sondes et des cibles qui, pour être évaluées, demandent de longs temps de calcul. La prise en compte d'un maximum de critères par le logiciel pourrait donc limiter le nombre de sondes nécessaires à la détection de chaque gène ciblé. Il faut cependant garder à l'esprit que la thermodynamique des hybridations des acides nucléiques reste mal connue, notamment s'agissant des réactions au niveau de l'interface liquide/solide (Pozhitkov *et al.* 2007). Malgré des caractéristiques thermodynamiques définies comme étant de bonne qualité, une sonde peut donc conduire à des résultats erronés (*e.g.* absence d'hybridation avec la cible ou hybridation aspécifique). Pour s'affranchir de cette limite, la stratégie actuelle consiste à sélectionner un groupe de sondes permettant de cibler différentes régions de chaque gène (Chou *et al.* 2004). Cependant, la mise en œuvre de cette stratégie est dépendante du gène ciblé et peut dans certains cas s'avérer difficile en raison, par exemple, de la taille et de la diversité des séquences au sein du groupe ciblé ou des critères (*e.g.* taille, paramètres thermodynamiques) auxquels doivent répondre chaque sonde. Une autre amélioration, liée à l'évolution constante des bases de données internationales comme GenBank (Benson *et al.* 2014), EMBL (Brooksbank *et al.* 2014) et DDBJ (Kosuge *et al.* 2014), serait de recalculer régulièrement les jeux de sondes afin d'avoir une meilleure estimation des paramètres de couverture et de spécificité. Malheureusement, la mauvaise qualité des bases de données internationales ne les rend pas directement exploitables (*i.e.* redondance, erreurs de séquençage, chimères (Ashelford *et al.*

2005)). Les récents développements de ProKSpOD tendent à intégrer la recherche, le contrôle qualité et l'organisation automatisés des séquences utilisées pour la détermination des sondes.

D'un point de vue moléculaire, nous avons pu montrer que la détermination de sondes de qualité pouvait permettre la mise en place d'une nouvelle approche d'exploration de la diversité microbienne à travers la capture de gènes. Cette approche particulièrement prometteuse a d'ores et déjà été appliquée à de nombreux autres biomarqueurs et écosystèmes. L'étude de la diversité taxonomique d'environnements aussi variés que des sols ou le microbiome d'un insecte a pu être effectuée en capturant spécifiquement les gènes codant pour la petite sous-unité du ribosome (*i.e.* 16S et 18S). Les capacités métaboliques ont également pu être étudiées grâce, par exemple, à la capture du biomarqueur de la méthanogénèse (*mcrA*) au sein d'un environnement lacustre, ou la capture d'enzymes fibrolytiques dans des échantillons de sols. En permettant de piéger de grands fragments d'ADN, la capture ainsi que l'émergence des technologies de séquençage de troisième génération, devrait permettre de séquencer en une seule fois de longues molécules d'ADN d'intérêt. Cette perspective intéressante rend alors envisageable la capture de fragments chromosomiques de grande taille voire de génomes complets. A l'inverse, cette méthode d'enrichissement peut être détournée pour procéder à la déplétion de certaines séquences dans des échantillons complexes. C'est notamment le cas lors d'études transcriptomiques ou métatranscriptomiques où le séquençage massif aboutit à la production de plus de 95% de séquences d'ADN ribosomiques. Des sondes de capture généralistes ciblant les gènes ADNr 16S, 18S, 23S et 28S ont donc été déterminées pour procéder à la déplétion. De plus, les protéines ribosomiques polluent également les jeux de données transcriptomiques et peuvent, elles aussi, être ciblées.

En résumé, les résultats obtenus au cours de cette thèse lient le développement d'outils bioinformatiques et moléculaires innovants à l'acquisition de données massives sur les communautés microbiennes d'environnements complexes. L'exploitation de ces données nécessite une complémentarité de compétences entre biologistes, bioinformaticiens et informaticiens. Cette philosophie pluridisciplinaire est indispensable pour relever les défis scientifiques et méthodologiques de la génomique environnementale à l'ère du *Big Data*.

Références

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **12**, 281–290.

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, **71**, 8966–8969.

Adessi C, Matton G, Ayala G *et al.* (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*, **28**, E87.

Ahmadian A, Ehn M, Hober S (2006) Pyrosequencing: history, biochemistry and future. *Clinica chimica acta ; international journal of clinical chemistry*, **363**, 83–94.

Ahn J, Yang L, Paster BJ *et al.* (2011) Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PLoS One*, **6**, e22788.

Allsopp D, Colwell RR, Hawksworth DL (1995) *Microbial diversity and ecosystem function*. CABI.

Alm EW, Oerther DB, Larsen N, Stahl DA, Raskin L (1996) The oligonucleotide probe database. *Applied and Environmental Microbiology*, **62**, 3557–3559.

Alonso-Alemany D, Barré A, Beretta S *et al.* (2013) Further Steps in TANGO: Improved Taxonomic Assignment in Metagenomics. *Bioinformatics (Oxford, England)*.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Amann RI, Krumholz L, Stahl DA (1990) Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *Journal of Bacteriology*, **172**, 762–770.

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, **59**, 143–169.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New biotechnology*, **25**, 195–203.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, **71**, 7724–7736.

Bachy C, Dolan JR, López-García P, Deschamps P, Moreira D (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *The ISME journal*, **7**, 244–255.

Bader KC, Grothoff C, Meier H (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics (Oxford, England)*, **27**, 1546–1554.

Bai S, Li J, He Z *et al.* (2013) GeoChip-based analysis of the functional gene diversity and metabolic potential of soil microbial communities of mangroves. *Applied Microbiology and Biotechnology*, **97**, 7035–7048.

Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, **13**, 92.

Beazley MJ, Martinez RJ, Rajan S *et al.* (2012) Microbial community analysis of a coastal salt marsh affected by the Deepwater Horizon oil spill. *PLoS One*, **7**, e41305.

Benson DA, Clark K, Karsch-Mizrachi I *et al.* (2014) GenBank. *Nucleic Acids Research*, **42**, D32–7.

Bentley DR (2006) Whole-genome re-sequencing. *Current opinion in genetics & development*, **16**, 545–552.
Berg RD (1996) The indigenous gastrointestinal microflora. *Trends in microbiology*, **4**, 430–435.
Berger SA, Stamatakis A (2011) Aligning short reads to reference alignments and trees. *Bioinformatics (Oxford, England)*, **27**, 2068–2075.
Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, **37**, 407–427.
Blochl E, Rachel R, Burggraf S *et al.* (1997) Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles*, **1**, 14–21.
Bodrossy L, Stralis-Pavese N, Murrell JC *et al.* (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental Microbiology*, **5**, 566–582.
Bontemps C, Golfier G, Gris-Liebe C *et al.* (2005) Microarray-based detection and typing of the Rhizobium nodulation gene nodC: potential of DNA arrays to diagnose biological functions of interest. *Applied and Environmental Microbiology*, **71**, 8042–8048.
Bottari B, Ercolini D, Gatti M, Neviani E (2006) Application of FISH technology for microbiological analysis: current state and prospects. *Applied Microbiology and Biotechnology*, **73**, 485–494.
Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, **6**, 673–676.
Brodie EL, DeSantis TZ, Joyner DC *et al.* (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Applied and Environmental Microbiology*, **72**, 6288–6298.
Brooksbank C, Bergman MT, Apweiler R, Birney E, Thornton J (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Research*, **42**, D18–25.
Call DR, Bakko MK, Krug MJ, Roberts MC (2003) Identifying antimicrobial resistance genes with DNA microarrays. *Antimicrobial agents and chemotherapy*, **47**, 3290–3295.
Capella-Gutierrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, **25**, 1972–1973.
Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.
Caporaso JG, Lauber CL, Walters WA *et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, **108 Suppl 1**, 4516–4522.
Case RJ, Boucher Y, Dahllöf I *et al.* (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, **73**, 278–288.
Chan C-KK, Hsu AL, Halgamuge SK, Tang S-L (2008) Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, **9**, 215.
Chan Y, Van Nostrand JD, Zhou J, Pointing SB, Farrell RL (2013) Functional ecology of an Antarctic Dry Valley. *Proceedings of the National Academy of Sciences*, **110**, 8990–8995.
Charuvaka A, Rangwala H (2011) Evaluation of short read metagenomic assembly. *BMC Genomics*, **12 Suppl 2**, S8.
Chatterjee S, Koslicki D, Dong S *et al.* (2014) SEK: Sparsity exploiting k-mer-based estimation of bacterial community composition. *Bioinformatics (Oxford, England)*, btu320.

Chatterji S, Yamazaki I, Bai Z, Eisen JA (2008) CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. *Proceedings of the 12th ….*

Cherf GM, Lieberman KR, Rashid H *et al.* (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, **30**, 344–348.

Chin C-S, Sorenson J, Harris JB *et al.* (2011) The origin of the Haitian cholera outbreak strain. *The New England journal of medicine*, **364**, 33–42.

Chou C-C, Chen C-H, Lee T-T, Peck K (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research*, **32**, e99.

Ciccarelli FD, Doerks T, Mering von C *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.

Claesson MJ, O'Sullivan O, Wang Q *et al.* (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One*, **4**, e6669.

Claesson MJ, Wang Q, O'Sullivan O *et al.* (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, **38**, e200.

Clarke J, Wu H-C, Jayasinghe L *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology*, **4**, 265–270.

Closek CJ, Sunagawa S, DeSalvo MK *et al.* (2014) Coral transcriptome and bacterial community profiles reveal distinct Yellow Band Disease states in Orbicella faveolata. *The ISME journal*.

Cole JR, Wang Q, Fish JA *et al.* (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, gkt1244.

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, **13**, 3021–3030.

Cowan DA, Russell NJ, Mamais A, Sheppard DM (2002) Antarctic Dry Valley mineral soils contain unexpectedly high levels of microbial biomass. *Extremophiles*, **6**, 431–436.

Cruz-Martínez K, Suttle KB, Brodie EL *et al.* (2009) Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *The ISME journal*, **3**, 738–744.

Darling AE, Jospin G, Lowe E *et al.* (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, **2**, e243.

Davenport CF, Neugebauer J, Beckmann N *et al.* (2012) Genometa - a fast and accurate classifier for short metagenomic shotgun reads. *PLoS One*, **7**, e41224.

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in bioinformatics*, **13**, 696–710.

Dean FB, Hosono S, Fang L *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5261–5266.

Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*, **11**, 1095–1099.

DeAngelis KM, Brodie EL, DeSantis TZ *et al.* (2009) Selective progressive response of soil microbial community to wild oat roots. *The ISME journal*.

Delmont TO, Robe P, Cecillon S *et al.* (2011) Accessing the soil metagenome for studies of microbial diversity. *Applied and Environmental Microbiology*, **77**, 1315–1324.

DeLong EF, Wickham GS, Pace NR (1989) Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, **243**, 1360–1363.

Denef VJ, Park J, Rodrigues JLM *et al.* (2003) Validation of a more sensitive method for using spotted oligonucleotide DNA microarrays for functional genomics studies on bacterial communities. *Environmental Microbiology*, **5**, 933–943.

Denonfoux J, Parisot N, Dugat-Bony E *et al.* (2013) Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **20**, 185–196.

Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F (2012) From genomics to metagenomics. *Current opinion in biotechnology*, **23**, 72–76.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, **10**, 56.

Ding G-C, Heuer H, He Z *et al.* (2012) More functional genes and convergent overall functional patterns detected by geochip in phenanthrene-spiked soils. *FEMS Microbiology Ecology*.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**, e105–e105.

Dröge J, McHardy AC (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in bioinformatics*, **13**, 646–655.

Dröge J, Gregor I, McHardy AC (2014) Taxator-tk: Fast and Precise Taxonomic Assignment of Metagenomes by Approximating Evolutionary Neighborhoods. *arXiv.org*.

Duc L, Neuenschwander S, Rehrauer H *et al.* (2009) Development and experimental validation of a nifH oligonucleotide microarray to study diazotrophic communities in a glacier forefield. *Environmental Microbiology*, **11**, 2179–2189.

Dufva M (2005) Fabrication of high quality microarrays. *Biomolecular engineering*, **22**, 173–184.

Dufva M (2009) Fabrication of DNA Microarray. *DNA Microarrays for Biomedical Research*, **529**, 63–79.

Dugat-Bony E, Biderre-Petit C, Jaziri F *et al.* (2012a) In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. *Microbial biotechnology*, **5**, 642–653.

Dugat-Bony E, Missaoui M, Peyretaillade E *et al.* (2011) HiSpOD: probe design for functional DNA microarrays. *Bioinformatics (Oxford, England)*, **27**, 641–648.

Dugat-Bony E, Peyretaillade E, Parisot N *et al.* (2012b) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environmental Microbiology*, **14**, 356–371.

Dunbar J, Barns SM, Ticknor LO, Kuske CR (2002) Empirical and theoretical bacterial diversity in four Arizona soils. *Applied and Environmental Microbiology*, **68**, 3035–3045.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*.

Edwards RA, Olson RJ, Disz T *et al.* (2012) Real Time Metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics (Oxford, England)*.

Edwards RA, Rodriguez-Brito B, Wegley L *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Ehrenreich A (2006) DNA microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology*, **73**, 255–273.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, **5**, e82.

Eisenstein M (2012) Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, **30**, 295–296.

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*, **34**, e22.

Finn RD, Bateman A, Clements J *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Research*, **42**, D222–30.

Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Applied and Environmental Microbiology*, **65**, 4630–4636.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**, 3150–3152.

Garmendia L, Hernandez A, Sanchez MB, Martinez JL (2012) Metagenomics and antibiotics. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, **18 Suppl 4**, 27–31.

Gelsomino A, Keijzer-Wolters AC, Cacco G, van Elsas JD (1999) Assessment of bacterial community structure in soil by polymerase chain reaction and denaturing gradient gel electrophoresis. *Journal of Microbiological Methods*, **38**, 1–15.

Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J (2006) Microarray applications in microbial ecology research. *Microbial Ecology*, **52**, 159–175.

Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, **39**, e91.

Ghebremedhin B, Layer F, König W, König B (2008) Genetic classification and distinguishing of Staphylococcus species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf gene sequences. *Journal of clinical microbiology*, **46**, 1019–1025.

Ghosh TS, Gajjalla P, Mohammed MH, Mande SS (2012) C16S - A Hidden Markov Model based algorithm for taxonomic classification of 16S rRNA gene sequences. *Genomics*.

Ghosh TS, Mohammed MH, Komanduri D, Mande SS (2011) ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation*, **6**, 91–94.

Ghosh TS, Monzoorul Haque M, Mande SS (2010) DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, **11 Suppl 7**, S14.

Gilles A, Meglécz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.

Gillespie DE, Brady SF, Bettermann AD *et al.* (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Applied and Environmental Microbiology*, **68**, 4301–4306.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources*, **11**, 759–769.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Gori F, Folino G, Jetten MSM, Marchiori E (2011) MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics (Oxford, England)*, **27**, 196–203.

Gouy M, Delmotte S (2008) Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie*, **90**, 555–562.

Guo J, Xu N, Li Z *et al.* (2008) Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences*, **105**, 9145–9150.

Guschin DY, Mobarry BK, Proudnikov D *et al.* (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Applied and Environmental Microbiology*, **63**, 2397–2402.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews : MMBR*, **68**, 669–685.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, **5**, R245–9.

Hartman AL, Riddle S, McPhillips T, Ludäscher B, Eisen JA (2010) Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. *BMC Bioinformatics*, **11**, 317.

He Z, Gentry TJ, Schadt CW *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME journal*, **1**, 67–77.

He Z, Piceno Y, Deng Y *et al.* (2012) The phylogenetic composition and structure of soil microbial communities shifts in response to elevated carbon dioxide. *The ISME journal*, **6**, 259–272.

He Z, Van Nostrand JD, Deng Y, Zhou J (2011) Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. *Frontiers of Environmental Science & Engineering in China*, **5**, 1–20.

Henne A, Schmitz RA, Bömeke M, Gottschalk G, Daniel R (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. *Applied and Environmental Microbiology*, **66**, 3113–3116.

Hess M, Sczyrba A, Egan R *et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463–467.

Hoff KJ (2009) The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, **10**, 520.

Horton M, Bodenhausen N, Bergelson J (2010) MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences. *Bioinformatics (Oxford, England)*, **26**, 568–569.

Huddleston J, Ranade S, Malig M *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, **24**, 688–696.

Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biology*, **3**, REVIEWS0003.

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, **180**, 4765–4774.

Hugoni M, Taib N, Debroas D *et al.* (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 6004–6009.

Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. *Science*, **254**, 59–67.

Hunter S, Corbett M, Denise H *et al.* (2013) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*.

Hunter S, Jones P, Mitchell A *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, **40**, D306–12.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Huyghe A, François P, Charbonnier Y *et al.* (2008) Novel microarray design strategy to study complex bacterial communities. *Applied and Environmental Microbiology*, **74**, 1876–1885.

Hyman ED (1988) A new method of sequencing DNA. *Analytical biochemistry*, **174**, 423–436.

Hysom DA, Naraghi-Arani P, Elsheikh M *et al.* (2012) Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. *PLoS One*, **7**, e34560.

Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS (2008) Genomic sequencing of single microbial cells from environmental samples. *Current Opinion in Microbiology*, **11**, 198–204.

Iwai S, Kurisu F, Urakawa H *et al.* (2008) Development of an oligonucleotide microarray to detect di- and monooxygenase genes for benzene degradation in soil. *FEMS microbiology letters*, **285**, 111–121.

Iwai S, Kurisu F, Urakawa H, Yagi O, Furumai H (2007) Development of a 60-mer oligonucleotide microarray on the basis of benzene monooxygenase gene diversity. *Applied Microbiology and Biotechnology*, **75**, 929–939.

Jacquiod S, Demanèche S, Franqueville L *et al.* (2014) Characterization of new bacterial catabolic genes and mobile genetic elements by high throughput genetic screening of a soil metagenomic library. *Journal of biotechnology*.

Jaing C, Gardner S, McLoughlin K *et al.* (2008) A functional gene array for detection of bacterial virulence elements. *PLoS One*, **3**, e2163.

Jaziri F, Parisot N, Abid A *et al.* (2014a) PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification. *Database : the Journal of Biological Databases and Curation*, **2014**, bau036–bau036.

Jaziri F, Peyretaillade E, Missaoui M *et al.* (2014b) Large Scale Explorative Oligonucleotide Probe Selection for Thousands of Genetic Groups on a Computing Grid: Application to Phylogenetic Probe Design Using a Curated Small Subunit Ribosomal RNA Gene Database. *The Scientific World Journal*, **2014**, 9–9.

Jiang B, Song K, Ren J *et al.* (2012) Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, **13**, 730.

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian protein metabolism* (ed Munro MN), pp. 21–132. Academic Press, N. Y.

Kafatos FC, Jones CW, Efstratiadis A (1979) Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Research*, **7**, 1541–1552.

Kanehisa M, Goto S, Sato Y *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, **42**, D199–205.

Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, **179**, 3899–3913.

Kawasaki ES (2006) The end of the microarray Tower of Babel: will universal standards lead the way?, **17**, 200–206.

Kelley DR, Salzberg SL (2010) Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics*, **11**, 544.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2011) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*.

Kelly JJ, Siripong S, McCormack J *et al.* (2005) DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. *Water Research*, **39**, 3229–3238.
Kelly LC, Cockell CS, Herrera-Belaroussi A *et al.* (2011) Bacterial Diversity of Terrestrial Crystalline Volcanic Rocks, Iceland. *Microbial Ecology*, **62**, 69–79.
Kelly LC, Cockell CS, Piceno YM *et al.* (2010) Bacterial Diversity of Weathered Terrestrial Icelandic Volcanic Glasses. *Microbial Ecology*, **60**, 740–752.
Kirk JL, Beaudette LA, Hart M *et al.* (2004) Methods of studying soil microbial diversity. *Journal of Microbiological Methods*, **58**, 169–188.
Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS (2009) Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, **10**, 316.
Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, **361**, 1929–1940.
Korbel JO, Jensen LJ, Mering von C, Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology*, **22**, 911–917.
Korlach J, Bjornson KP, Chaudhuri BP *et al.* (2010) Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology*, **472**, 431–455.
Korves TM, Piceno YM, Tom LM *et al.* (2013) Bacterial communities in commercial aircraft high-efficiency particulate air (HEPA) filters assessed by PhyloChip analysis. *Indoor Air*, **23**, 50–61.
Koslicki D, Foucart S, Rosen GL (2014) WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification. (MR Liles, Ed,). *PLoS One*, **9**, e91784.
Kosuge T, Mashima J, Kodama Y *et al.* (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Research*, **42**, D44–9.
Kotamarti RM, Hahsler M, Raiford D, McGee M, Dunham MH (2010) Analyzing taxonomic classification using extensible Markov models. *Bioinformatics (Oxford, England)*, **26**, 2235–2241.
Kubota K (2013) CARD-FISH for environmental microorganisms: technical advancement and future applications. *Microbes and environments / JSME*, **28**, 3–12.
Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.
Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, **23**, 2947–2948.
Lasken RS (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology*, **10**, 510–516.
Lasken RS (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, **10**, 631–640.
Laszlo AH, Derrington IM, Ross BC *et al.* (2014) Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*.
Leach ALB, Chong JPJ, Redeker KR (2012) SSuMMo: rapid analysis, comparison and visualization of microbial communities. *Bioinformatics (Oxford, England)*, **28**, 679–686.
Lee DH, Zo YG, Kim SJ (1996) Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. *Applied and Environmental Microbiology*, **62**, 3112–3120.
Lee N, Nielsen PH, Andreasen KH *et al.* (1999) Combination of fluorescent in situ hybridization and microautoradiography-a new tool for structure-function analyses in microbial ecology. *Applied and Environmental Microbiology*, **65**, 1289–1297.

Lee PKH, Warnecke F, Brodie EL *et al.* (2012) Phylogenetic microarray analysis of a microbial community performing reductive dechlorination at a TCE-contaminated site. *Environmental Science & Technology*, **46**, 1044–1054.
Lee Y-J, Van Nostrand JD, Tu Q *et al.* (2013) The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. *The ISME journal*, –.
Leis B, Angelov A, Liebl W (2013) Screening and expression of genes from metagenomes. *Advances in applied microbiology*, **83**, 1–68.
Lemoine S, Combes F, Le Crom S (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Research*, **37**, 1726–1739.
Levene MJ, Korlach J, Turner SW *et al.* (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.
Li H, Groep D, Wolters L, Templon J (2006) Job Failure Analysis and Its Implications in a Large-Scale Production Grid. In:, pp. 27–27. IEEE.
Li K-B (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics (Oxford, England)*, **19**, 1585–1586.
Li T, Wu T-D, Mazéas L *et al.* (2008) Simultaneous analysis of microbial identity and function using NanoSIMS. *Environmental Microbiology*, **10**, 580–588.
Liles MR, Turkmen O, Manske BF *et al.* (2010) A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. *Soil biology & biochemistry*, **42**, 739–747.
Liu B, Gibbons T, Ghodsi M, Pop M (2010) MetaPhyler: Taxonomic profiling for metagenomic sequences. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 95–100.
Liu J, Wang H, Yang H *et al.* (2012a) Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Research*.
Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology*, **63**, 4516–4522.
Liu W, Li L, Khan MA, Zhu F (2012b) Popular molecular markers in bacteria. *Molekuliarnaia genetika, mikrobiologiia i virusologiia*, 14–17.
Logares R, Haverkamp THA, Kumar S *et al.* (2012) Environmental microbiology through the lens of High-Throughput DNA Sequencing: Synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods*.
Loy A, Bodrossy L (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clinica chimica acta ; international journal of clinical chemistry*, **363**, 106–119.
Loy A, Lehner A, Lee N *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Applied and Environmental Microbiology*, **68**, 5064–5081.
Loy A, Maixner F, Wagner M, Horn M (2007) probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Research*, **35**, D800–4.
Loy A, Schulz C, Lücker S *et al.* (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "Rhodocyclales". *Applied and Environmental Microbiology*, **71**, 1373–1386.
Luke C, Frenzel P (2011) Potential of pmoA Amplicon Pyrosequencing for Methanotroph Diversity Studies. *Applied and Environmental Microbiology*, **77**, 6305–6309.
Luo C, Rodriguez-R LM, Konstantinidis KT (2014) MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, gku169.

Macdonald NJ, Parks DH, Beiko RG (2012) Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Research*.

Mande SS, Mohammed MH, Ghosh TS (2012) Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*.

Manrao EA, Derrington IM, Laszlo AH *et al.* (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, **30**, 349–353.

Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, **27**, 764–770.

Marcelino LA, Backman V, Donaldson A *et al.* (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13629–13634.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Martin C, Diaz NN, Ontrup J, Nattkemper TW (2008) Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics (Oxford, England)*, **24**, 1568–1574.

Martín R, Miquel S, Langella P, Bermúdez-Humarán LG (2014) The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, **5**, 413–423.

Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

McCarthy A (2010) Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & Biology*, **17**, 675–676.

McDonald D, Price MN, Goodrich JK *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, **6**, 610–618.

McHardy AC, Rigoutsos I (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. *Current Opinion in Microbiology*, **10**, 499–503.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, **4**, 63–72.

McLean JS, Lombardo M-J, Badger JH *et al.* (2013) Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences*, **110**, E2390–9.

McNally B, Singer A, Yu Z *et al.* (2010) Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano letters*, **10**, 2237–2244.

Meacham F, Boffelli D, Dhahbi J *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Meinicke P, Asshauer KP, Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics (Oxford, England)*, **27**, 1618–1624.

Mellmann A, Harmsen D, Cummings CA *et al.* (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, **6**, e22751.

Mering von C, Hugenholtz P, Raes J *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.

Metzker ML (2005) Emerging technologies in DNA sequencing. *Genome Research*, **15**, 1767–1776.

Metzker ML (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**, 31–46.
Meyer F, Paarmann D, D'Souza M *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
Militon C, Rimour S, Missaoui M *et al.* (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics (Oxford, England)*, **23**, 2550–2557.
Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.
Miller SM, Tourlousse DM, Stedtfeld RD *et al.* (2008) In situ-synthesized virulence and marker gene biochip for detection of bacterial pathogens in water. *Applied and Environmental Microbiology*, **74**, 2200–2209.
Mitra S, Rupek P, Richter DC *et al.* (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, **12 Suppl 1**, S21.
Mohammed MH, Chadaram S, Komanduri D, Ghosh TS, Mande SS (2011a) Eu-Detect: An algorithm for detecting eukaryotic sequences in metagenomic data sets. *Journal of biosciences*, **36**, 709–717.
Mohammed MH, Ghosh TS, Reddy RM *et al.* (2011b) INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics*, **12 Suppl 3**, S4.
Mohammed MH, Ghosh TS, Singh NK, Mande SS (2011c) SPHINX--an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics (Oxford, England)*, **27**, 22–30.
Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS (2009) SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics (Oxford, England)*, **25**, 1722–1730.
Morales SE, Holben WE (2011) Linking bacterial identities and ecosystem processes: can “omic” analyses be more than the sum of their parts? *FEMS Microbiology Ecology*, **75**, 2–16.
Morey M, Fernández-Marmiesse A, Castiñeiras D *et al.* (2013) A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism*.
Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.
Munroe DJ, Harris TJR (2010) Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology*, **28**, 426–428.
Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*, **59**, 695–700.
Nagarajan N, Pop M (2013) Sequence assembly demystified. *Nature Reviews Genetics*, **14**, 157–167.
Nakamura K, Oshima T, Morimoto T *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, **39**, e90.
Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K (2011) RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, **12**, 41.
Narihiro T, Sekiguchi Y (2011) Oligonucleotide primers, probes and molecular methods for the environmental monitoring of methanogenic archaea. *Microbial biotechnology*, **4**, 585–602.

Nawy T (2014) Single-cell sequencing. *Nature Methods*, **11**, 18.
NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **42**, D7–17.
Nemir A, David MM, Perrussel R *et al.* (2010) Comparative phylogenetic microarray analysis of microbial communities in TCE-contaminated soils. *Chemosphere*, **80**, 600–607.
Niu B, Zhu Z, Fu L, Wu S, Li W (2011) FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics (Oxford, England)*, **27**, 1704–1705.
Nyrén P, Lundin A (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical biochemistry*, **151**, 504–509.
Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–2901.
Owens B (2012) Genomics: The single life. *Nature*, **491**, 27–29.
Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.
Paliy O, Agans R (2012) Application of phylogenetic microarrays to interrogation of human microbiota. *FEMS Microbiology Ecology*, **79**, 2–11.
Palmer C, Bik EM, Eisen MB *et al.* (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Research*, **34**, e5.
Pandit AS, Joshi MN, Bhargava P *et al.* (2014) Metagenomes from the saline desert of kutch. *Genome announcements*, **2**, e00439–14.
Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *Journal of applied genetics*, **52**, 413–435.
Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E (2012) KASpOD--a web service for highly specific and explorative oligonucleotide design. *Bioinformatics (Oxford, England)*, **28**, 3161–3162.
Parisot N, Denonfoux J, Dugat-Bony E, Peyretaillade E, Peyret P (2014) Software Tools for the Selection of Oligonucleotide Probes for Microarrays. In: *Microarrays: Current Technology, Innovations and Applications* (ed He Z), p. 250. Academic Press.
Parks DH, Macdonald NJ, Beiko RG (2011) Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, **12**, 328.
Pathak A, Shanker R, Garg SK, Manickam N (2011) Profiling of biodegradation and bacterial 16S rRNA genes in diverse contaminated ecosystems using 60-mer oligonucleotide microarray. *Applied Microbiology and Biotechnology*, **90**, 1739–1754.
Pati A, Heath LS, Kyrpides NC, Ivanova NN (2011) ClaMS: A Classifier for Metagenomic Sequences. *Standards in genomic sciences*, **5**, 248–253.
Patil KR, Roune L, McHardy AC (2012) The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS One*, **7**, e38581.
Patin NV, Kunin V, Lidström U, Ashby MN (2012) Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. *Microbial Ecology*.
Pedrós-Alió C (2007) Ecology. Dipping into the rare biosphere. *Science*, **315**, 192–193.
Pedrós-Alió C (2012) The rare bacterial biosphere. *Annual review of marine science*, **4**, 449–466.
Pelletier E, Perrière G (2013) Accès et partage des données NGS. In: *Les cahiers prospectives Génomique environnementale (INEE-CNRS)* (ed CNRS). CNRS.
Peplies J, Lau SCK, Pernthaler J, Amann RL, Glöckner FO (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environmental Microbiology*, **6**, 638–645.

Petrosino JF, Highlander SK, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. *Clinical chemistry*, **55**, 856–866.
Peyret P (2013) Qualité des données NGS : de la séquence aux bases de données. In: *Les cahiers prospectives Génomique environnementale (INEE-CNRS)* (ed CNRS). CNRS.
Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.
Porter MS, Beiko RG (2013) SPANNER: taxonomic assignment of sequences using pyramid matching of similarity profiles. *Bioinformatics (Oxford, England)*, **29**, 1858–1864.
Pozhitkov AE, Tautz D, Noble PA (2007) Oligonucleotide microarrays: widely applied--poorly understood. *Briefings in functional genomics & proteomics*, **6**, 141–148.
Prabhakara S, Acharya R (2010) SIMCOMP: A Hybrid Soft Clustering of Metagenome Reads. *Pattern Recognition in Bioinformatics*.
Prakash T, Taylor TD (2012) Functional assignment of metagenomic data: challenges and applications. *Briefings in bioinformatics*.
Preza D, Olsen I, Willumsen T *et al.* (2009) Microarray analysis of the microflora of root caries in elderly. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, **28**, 509–517.
Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
Prüfer K, Stenzel U, Dannemann M *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics (Oxford, England)*, **24**, 1530–1531.
Qin J, Li R, Raes J *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
Qu W, Hashimoto S-I, Morishita S (2009) Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Research*, **19**, 1309–1315.
Quail MA, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–6.
Quince C, Curtis TP, Sloan WT (2008) The rational exploration of microbial diversity. *The ISME journal*, **2**, 997–1006.
Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.
Rachamalla MR, Monzoorul Haque M, Mande SS (2012) TWARIT: An extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene*.
Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature*, **403**, 646–649.
Raghunathan A, Ferguson HR, Bornarth CJ *et al.* (2005) Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology*, **71**, 3342–3347.
Rajendhran J, Gunasekaran P (2008) Strategies for accessing soil metagenome for desired applications. *Biotechnology Advances*, **26**, 576–590.
Rajilić-Stojanović M, Heilig HGHJ, Molenaar D *et al.* (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environmental Microbiology*, **11**, 1736–1751.

Rajilić-Stojanović M, Heilig HGHJ, Tims S, Zoetendal EG, de Vos WM (2012) Long-term monitoring of the human intestinal microbiota composition. *Environmental Microbiology*, **15**, 1146–1159.

Ramette A (2009) Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. *Applied and Environmental Microbiology*, **75**, 2495–2505.

Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annual review of microbiology*, **57**, 369–394.

Rastogi G, Barua S, Sani RK, Peyton BM (2011) Investigation of Microbial Populations in the Extremely Metal-Contaminated Coeur d'Alene River Sediments. *Microbial Ecology*, **62**, 1–13.

Relógio A, Schwager C, Richter A, Ansorge W, Valcárcel J (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research*, **30**, e51.

Rhee S-K, Liu X, Wu L *et al.* (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Applied and Environmental Microbiology*, **70**, 4303–4317.

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, **38**, e191.

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annual review of genetics*, **38**, 525–552.

Rimour S, Hill DRC, Militon C, Peyret P (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics (Oxford, England)*, **21**, 1094–1103.

Rinke C, Schwientek P, Sczyrba A *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*.

Rinta-Kanto JM, Bürgmann H, Gifford SM *et al.* (2011) Analysis of sulfur-related transcription by Roseobacter communities using a taxon-specific functional gene microarray. *Environmental Microbiology*, **13**, 453–467.

Roder C, Arif C, Bayer T *et al.* (2014) Bacterial profiling of White Plague Disease in a comparative coral species framework. *The ISME journal*, **8**, 31–39.

Rodrigue S, Materna AC, Timberlake SC *et al.* (2010) Unlocking short read sequencing for metagenomics. *PLoS One*, **5**, e11840.

Roh SW, Abell GCJ, Kim K-H, Nam Y-D, Bae J-W (2010) Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in Biotechnology*, **28**, 291–299.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, **242**, 84–89.

Rose TM, Henikoff JG, Henikoff S (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Research*, **31**, 3763–3766.

Rose TM, Schultz ER, Henikoff JG *et al.* (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Research*, **26**, 1628–1635.

Rosen GL, Reichenberger ER, Rosenfeld AM (2011) NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics (Oxford, England)*, **27**, 127–129.

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nature Biotechnology*, **26**, 1117–1124.

Rothberg JM, Hinz W, Rearick TM *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Rusch DB, Halpern AL, Sutton G *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **5**, e77.
Saiki RK, Gelfand DH, Stoffel S *et al.* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
Sait M, Hugenholtz P, Janssen PH (2002) Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environmental Microbiology*, **4**, 654–666.
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.
Sanguin H, Sarniguet A, Gazengel K, Moënne-Loccoz Y, Grundmann GL (2009) Rhizosphere bacterial communities associated with disease suppressiveness stages of take-all decline in wheat monoculture. *The New phytologist*, **184**, 694–707.
Santos SR, Ochman H (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology*, **6**, 754–759.
Sapkota AR, Berger S, Vogel TM (2010) Human pathogens abundant in the bacterial metagenome of cigarettes. *Environmental health perspectives*, **118**, 351–356.
Savage DC (1977) Microbial ecology of the gastrointestinal tract. *Annual review of microbiology*, **31**, 107–133.
Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010a) Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, **11**, 647–657.
Schadt EE, Turner S, Kasarskis A (2010b) A window into third-generation sequencing. *Human molecular genetics*, **19**, R227–40.
Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. *Genome Research*, **20**, 1165–1173.
Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
Schloss PD, Handelsman J (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, **9**, 34.
Schloss PD, Westcott SL (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology*, **77**, 3219–3226.
Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, **6**, e27310.
Schneider GF, Dekker C (2012) DNA sequencing with nanopores. *Nature Biotechnology*, **30**, 326–328.
Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, **5**, e1000605.
Scholten JCM, Culley DE, Nie L *et al.* (2007) Development and assessment of whole-genome oligonucleotide microarrays to analyze an anaerobic microbial community and its responses to oxidative stress. *Biochemical and biophysical research communications*, **358**, 571–577.
Schönhuber W, Fuchs B, Juretschko S, Amann R (1997) Improved sensitivity of whole-cell hybridization by the combination of horseradish peroxidase-labeled oligonucleotides and tyramide signal amplification. *Applied and Environmental Microbiology*, **63**, 3268–3273.

Schönmann S, Loy A, Wimmersberger C *et al.* (2009) 16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genus Burkholderia. *Environmental Microbiology*, **11**, 779–800.
Schreiber F, Gumrich P, Daniel R, Meinicke P (2010) Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics (Oxford, England)*, **26**, 960–961.
Segata N, Waldron L, Ballarini A *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*.
Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. *PLoS Biology*, **5**, e75.
Sharma VK, Kumar N, Prakash T, Taylor TD (2012) Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin. *PLoS One*, **7**, e34030.
Sharpton TJ, Riesenfeld SJ, Kembel SW *et al.* (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology*, **7**, e1001061.
Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.
Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular ecology*, **21**, 1794–1805.
Shumway M, Cochrane G, Sugawara H (2010) Archiving next generation sequencing data. *Nucleic Acids Research*, **38**, D870–1.
Sigrist CJA, de Castro E, Cerutti L *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Research*, **41**, D344–7.
Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, **77**, 1153–1161.
Singer A, McNally B, Torre RD, Meller A (2012) DNA sequencing by nanopore-induced photon emission. *Methods in molecular biology (Clifton, N.J.)*, **870**, 99–114.
Smith DJ, Timonen HJ, Jaffe DA *et al.* (2013) Intercontinental dispersal of bacteria and archaea by transpacific winds. *Applied and Environmental Microbiology*, **79**, 1134–1139.
Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.
Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, **98**, 503–517.
Srinivasan SM, Guda C (2013) MetaID: A novel method for identification and quantification of metagenomic samples. *BMC Genomics*, **14 Suppl 8**, S4.
Stark M, Berger SA, Stamatakis A, Mering von C (2010) MLTreeMap--accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics*, **11**, 461.
Steele HL, Jaeger K-E, Daniel R, Streit WR (2009) Advances in recovery of novel biocatalysts from metagenomes. *Journal of Molecular Microbiology and Biotechnology*, **16**, 25–37.
Stralis-Pavese N, Abell GCJ, Sessitsch A, Bodrossy L (2011) Analysis of methanotroph community composition using a pmoA-based microbial diagnostic microarray. *Nature Protocols*, **6**, 609–624.
Stranneheim H, Käller M, Allander T *et al.* (2010) Classification of DNA sequences using Bloom filters. *Bioinformatics (Oxford, England)*, **26**, 1595–1600.
Su X, Pan W, Song B, Xu J, Ning K (2014) Parallel-META 2.0: Enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization. *PLoS One*, **9**, e89323.

Suenaga H (2011) Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environmental Microbiology*.
Sunagawa S, Mende DR, Zeller G *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*.
Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.
Swerdlow H, Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, **18**, 1415–1419.
Taib N, Mangot J-F, Domaizon I, Bronner G, Debroas D (2013) Phylogenetic Affiliation of SSU rRNA Genes Generated by Massively Parallel Sequencing: New Insights into the Freshwater Protist Diversity. *PLoS One*, **8**, e58950.
Tanaseichuk O, Borneman J, Jiang T (2013) Phylogeny-based classification of microbial communities. *Bioinformatics (Oxford, England)*, btt700.
Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Applied and Environmental Microbiology*, **69**, 1159–1171.
Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
Teeling H, Waldmann J, Lombardot T, Bauer M, Gloeckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
Terrat S, Peyretaillade E, Goncalves O *et al.* (2010) Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics*, **11**, 478.
Timp W, Mirsaidov UM, Wang D *et al.* (2010) Nanopore Sequencing: Electrical Measurements of the Code of Life. *IEEE transactions on nanotechnology*, **9**, 281–294.
Tiquia SM, Gurczynski S, Zholi A, Devol A (2006) Diversity of biogeochemical cycling genes from Puget Sound sediments using DNA microarrays. *Environmental technology*, **27**, 1377–1389.
Torsvik V, Goksøyr J, Daae FL (1990) High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology*, **56**, 782–787.
Tottey W, Denonfoux J, Jaziri F *et al.* (2013) The Human Gut Chip "HuGChip," an Explorative Phylogenetic Microarray for Determining Gut Microbiome Diversity at Family Level. *PLoS One*, **8**, e62544.
Tu Q, He Z, Zhou J (2014a) Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Research*, gku138.
Tu Q, He Z, Li Y *et al.* (2014b) Development of HuMiChip for Functional Profiling of Human Microbiomes. *PLoS One*, **9**, e90546.
Tu Q, Yu H, He Z *et al.* (2014c) GeoChip 4: a functional gene array-based high throughput environmental technology for microbial community analysis. *Molecular ecology resources*, n/a–n/a.
Turcatti G, Romieu A, Fedurco M, Tairi A-P (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, **36**, e25–e25.
Tuzhikov A, Panchin A, Shestopalov VI (2014) TUIT, a BLAST-based tool for taxonomic classification of nucleotide sequences. *BioTechniques*, **56**, 78–84.
Tyson GW, Chapman J, Hugenholtz P *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Valm AM, Mark Welch JL, Borisy GG (2012) CLASI-FISH: principles of combinatorial labeling and spectral imaging. *Systematic and applied microbiology*, **35**, 496–502.
Valm AM, Mark Welch JL, Rieken CW *et al.* (2011) Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proceedings of the National Academy of Sciences*, **108**, 4152–4157.
van den Bogert B, de Vos WM, Zoetendal EG, Kleerebezem M (2011) Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. *Applied and Environmental Microbiology*, **77**, 2071–2080.
Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
Vieites JM, Guazzaroni M-E, Beloqui A, Golyshin PN, Ferrer M (2009) Metagenomics approaches in systems microbiology. *FEMS Microbiology Reviews*, **33**, 236–255.
Wagner M, Nielsen PH, Loy A, Nielsen JL, Daims H (2006) Linking microbial community structure with function: fluorescence in situ hybridization-microautoradiography and isotope arrays. *Current opinion in biotechnology*, **17**, 83–91.
Wagner M, Smidt H, Loy A, Zhou J (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microbial Ecology*, **53**, 498–506.
Wang GY, Graziani E, Waters B *et al.* (2000) Novel natural products from soil DNA libraries in a streptomycete host. *Organic letters*, **2**, 2401–2404.
Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.
Wang X, Xia Y, Wen X, Yang Y, Zhou J (2014a) Microbial community functional structures in wastewater treatment plants as characterized by GeoChip. (Z Zhou, Ed,). *PLoS One*, **9**, e93422.
Wang Y, Leung H, Yiu S, Chin F (2014b) MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics*, **15 Suppl 1**, S12.
Ward BB, Bouskill NJ (2011) The utility of functional gene arrays for assessing community composition, relative abundance, and distribution of ammonia-oxidizing bacteria and archaea. *Methods in Enzymology*, **496**, 373–396.
Ward BB, Eveillard D, Kirshtein JD *et al.* (2007) Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology*, **9**, 2522–2538.
Wash S, Image C (2008) DNA sequencing: generation next–next. *Nature Methods*.
Weber M, Teeling H, Huang S *et al.* (2011) Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *The ISME journal*, **5**, 918–928.
Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6578–6583.
Wintzingerode von F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, **21**, 213–229.
Woese CR (1987) Bacterial evolution. *Microbiological Reviews*, **51**, 221–271.
Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4576–4579.

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. *Applied and Environmental Microbiology*, **74**, 1453–1463.
Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**, R46.
Wooley JC, Ye Y (2009) Metagenomics: Facts and Artifacts, and Computational Challenges. *Journal of computer science and technology*, **25**, 71–81.
Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. *PLoS Computational Biology*, **6**, e1000667.
Wu D, Hartman A, Ward N, Eisen JA (2008) An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP). *PLoS One*, **3**, e2566.
Wu L, Thompson DK, Li G *et al.* (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology*, **67**, 5780–5790.
Wu L, Thompson DK, Liu X *et al.* (2004) Development and evaluation of microarray-based whole-genome hybridization for detection of microorganisms within the context of environmental applications. *Environmental Science & Technology*, **38**, 6775–6782.
Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, **9**, R151.
Wu Y-W, Ye Y (2011) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, **18**, 523–534.
Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, **6**, e27992.
Yang Y, Gao Y, Wang S *et al.* (2014) The microbial gene diversity along an elevation gradient of the Tibetan grassland. *The ISME journal*, **8**, 430–440.
Yergeau E, Schoondermark-Stolk SA, Brodie EL *et al.* (2009) Environmental microarray analyses of Antarctic soil microbial communities. *The ISME journal*, **3**, 340–351.
Yoccoz NG (2012) The future of environmental DNA in ecology. *Molecular ecology*, **21**, 2031–2038.
Yu F, Sun Y, Liu L, Farmerie W (2010) GSTaxClassifier: a genomic signature based taxonomic classifier for metagenomic data analysis. *Bioinformation*, **4**, 46–49.
Yu K, Zhang T (2012) Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS One*, **7**, e38183.
Zakrzewski M, Bekel T, Ander C *et al.* (2012) MetaSAMS-A novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets. *Journal of biotechnology*.
Zhang T, Fang HHP (2006) Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Applied Microbiology and Biotechnology*, **70**, 281–289.
Zhao G, Bu D, Liu C *et al.* (2012) CloudLCA: finding the lowest common ancestor in metagenome analysis using cloud computing. *Protein & Cell*, **3**, 148–152.
Zhou J (2003) Microarrays for bacterial detection and microbial community analysis. *Current Opinion in Microbiology*, **6**, 288–294.
Zhou J, Thompson DK (2002) Challenges in applying microarrays to environmental studies. *Current opinion in biotechnology*, **13**, 204–207.
Zhou X, Ren L, Li Y *et al.* (2010) The next-generation sequencing technology: a technology review and future perspective. *Protein & Cell*, **53**, 44–57.
Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, **38**, e132.

Annexes

Solution hybrid selection capture for the recovery of functional full-length eukaryotic cDNAs from complex environmental samples

Authors

Claudia Bragalini[1,2], Céline Ribière[3], Nicolas Parisot[3], Laurent Vallon[2], Elsa Prudent[2], Eric Peyretaillade[3], Mariangela Girlanda[2,4], Pierre Peyret[3], Roland Marmeisse[1,2], Patricia Luis[2]

Affiliations

[1] Department of Life Sciences and Systems Biology, University of Turin, viale Mattioli 25, 10125 Turin, Italy

[2] Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université de Lyon, Université Lyon 1, 69622 Villeurbanne, France

[3] Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, F-63001 Clermont-Ferrand, France

[4] Istituto per la Protezione Sostenibile delle Piante (IPSP), Consiglio Nazionale delle Ricerche, Viale Mattioli 25, 10125 Turin, Italy

Author for correspondence

Patricia Luis

Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Bâtiment André Lwoff, 43 Boulevard du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France.

Tel: + 33 (0)472431050

E-mail: patricia.luis@univ-lyon1.fr

Abstract

Eukaryotic microbial communities play key functional roles in soil biology and potentially represent a rich source of natural products including biocatalysts. Culture-independent molecular methods are powerful tools to isolate functional genes from uncultured microorganisms. However, none of the methods used in environmental genomics allow for a rapid isolation of numerous functional genes from eukaryotic microbial communities. We developed an original adaptation of the solution hybrid selection (SHS) for an efficient recovery of functional cDNAs synthesized from soil-extracted polyadenylated mRNA. This protocol was tested on the Glycoside Hydrolase 11 gene family encoding endoxylanases for which we designed 35 explorative 31-mers capture probes. SHS was implemented on four soil eukaryotic cDNA pools. After two successive rounds of capture, more than 90% of the resulting cDNAs were GH11 sequences, of which 70% (38 among 53 sequenced genes) were full-length. Between 1.5 and 25% of the cloned captured sequences were expressed in *Saccharomyces cerevisiae*. Sequencing of PCR-amplified GH11 gene fragments from the captured sequences highlighted hundreds of phylogenetically diverse sequences that were not yet described in public databases. This protocol offers the possibility of performing exhaustive exploration of eukaryotic gene families within microbial communities thriving in any type of environment.

Keywords: metatranscriptomics; soil RNA; soil eukaryotes; sequence capture; glycoside hydrolase family GH11

1 Introduction

A common objective of many studies in the field of environmental microbiology is to evaluate the functional diversity of the complex microbial communities colonising natural or man-made environments, either fresh or marine waters, sediments, soils, digestive tracts or food products. This diversity can be apprehended through the systematic sequencing and functional annotation of DNA (metagenomics) or RNA (metatranscriptomics) molecules directly extracted from environmental samples[1,2]. However, as a result of the extreme taxonomic richness of most microbial communities, high-throughput shotgun sequencing of environmental nucleic acids is far from covering their full gene repertoire[3]. Alternatively, many studies focus on specific environmental processes which, for some of them, are controlled by a limited and defined set of genes encoding key enzymes. The diversity of the corresponding gene families and of the organisms that possess and express them is classically evaluated by the systematic sequencing and taxonomic annotation of PCR-amplified gene fragments from environmental DNA or RNA (metabarcoding)[4-7]. This latter approach has itself well documented limitations. One of the limitations is that the use of a single pair of degenerate primers, designed to hybridize to internal gene consensus sequences, usually fails to amplify all homologous sequences present in an environmental sample[8]. Another, often underestimated limitation is that metabarcoding does not allow amplification of full-length functional genes. Besides limiting the number of phylogenetically-informative nucleotide positions for precise phylogenetic assignment of environmental sequences, obtaining partial sequences also prevents their functional study by expression in a heterologous microbial host. Full-length functional genes are yet of importance (i) in ecology to establish potential relationships between enzyme catalytic properties (substrate range, sensitivity to physicochemical parameters) and prevailing environmental conditions and (ii) in environmental biotechnology to isolate novel biocatalysts for industrial purpose.

In a recent publication, Denonfoux *et al.*[9] developed an alternative strategy to explore microbial communities from complex environments. Based on solution hybrid selection (hereafter referred to

as SHS), this method allows for the specific recovery of large DNA fragments harbouring biomarkers of interest even from rare or unknown microorganisms. Indeed, SHS is based on the design of several oligonucleotide probes which can cover the whole gene of interest as opposed to PCR strategies targeting internal regions. Moreover, explorative probe design strategies using appropriate software such as HiSpOD[10] or KASpOD[11] allow recovering not yet described homologous sequences[9]. These probes are synthesized as biotinylated RNA oligonucleotides and hybridized, in solution, to the target gene sequences diluted among a majority of non-target DNA fragments. The hybrid molecules [biotinylated probes + target sequences] are then specifically captured by affinity-binding on streptavidin-coated paramagnetic beads. SHS can be repeated several times successively to increase the enrichment in desired sequences by a factor of up to 1.7 10^5 times[9]. In environmental microbiology the captured DNA fragments can be subjected to high-throughput sequencing. *In silico* assembly of the reads not only leads to the reconstruction of the full-length sequences of the different members of the targeted gene family but also of their genomic environment and could therefore facilitate operon reconstructions[9].

In microbial ecology, SHS has thus far been successfully used to capture archaeal protein-coding genes from environmental DNA[9]. As previously discussed[12], environmental DNA is however not the most appropriate matrix to recover full-length functional genes of eukaryotic origin, which could be easily expressed in a heterologous microbial host. Environmental polyadenylated messenger RNA, devoid of introns, represent a better source of eukaryotic genes which, following their conversion into cDNA, can be expressed in either bacteria or yeasts[12-16].

Soil eukaryotes such as fungi, are highly diverse[17,18], play essential roles in soil biology as for example the main agents in plant organic matter degradation[19,20] and represent a rich source of enzymes and biomolecules used in industry[21]. Despite these obvious interests, very few environmental genomics studies specifically focus on soil eukaryote functional diversity[22].

To promote such studies, we developed and evaluated in the present report an original adaptation of the SHS for the efficient recovery of full-length functional fungal cDNAs synthesized

from soil RNA. Successful development of this technique was favoured by the ever increasing number of available fungal genomes that provide a correspondingly large number of members of specific gene families for the design of hybridization probes[23]. The fungal gene family targeted in the present study is the Glycoside Hydrolase 11 (GH11) family which encode endo-ß-1,4 xylanases (E.C. 3.2.1.8) (CAZY Carbohydrate Active Enzymes database, http://www.cazy.org)[24]. These enzymes have an obvious importance in soil ecology for the degradation of plant hemicelluloses and are also abundantly used in different industrial processes[26]. GH11 genes are present in the genomes of numerous fungi, mainly Ascomycota and Basidiomycota, and at the start of this study, more than 300 sequences were publicly available. Furthermore, in a random shotgun sequencing of forest soil eukaryotic polyA-mRNAs, it was shown that GH11 transcripts occurred at a low frequency ranging from 0 to 1 per 10^4 sequences obtained[22,25].

2. Materials and methods

2.1. Soil RNA extraction and cDNA synthesis.

Four different forest soils from France and Italy were used in this study (Table S1). At each site, between 30 (BEW) and 60 (BRH) sieved (2 mm) soil cores were mixed together to constitute composite samples which were stored at -75°C prior to RNA extraction. RNA was extracted from 4 to 48 g of soil using protocols adapted to each soil. RNA from the Puéchabon (PUE) sample was extracted according to Luis *et al.*[5]. RNA from the Breuil Spruce (BRE) and Breuil Beech (BRH) samples were extracted according to Damon *et al.*[27]. RNA from the Berchidda (BEW) sample was extracted using the PowerSoil® Total RNA Isolation Kit (Mo Bio Laboratories), according to the manufacturer's instructions. All RNA samples were treated with RNase free DNase I, to remove residual DNA contaminations and quantified by spectrophotometry (ND-1000 NanoDrop®, Thermo Scientific).

123 Eukaryotic cDNAs were synthesized from 2 µg of total soil RNA by using the Mint-2 cDNA
124 synthesis and amplification kit according to the manufacturer's instructions (Evrogen). First strand
125 synthesis was initiated at the RNA 3' poly-A end using a modified poly-dT primer (CDS-4M). The
126 number of PCR cycles (between 22 and 30) necessary for optimal synthesis of the double stranded
127 cDNA (dscDNA) was evaluated for each cDNA sample. As a result of using the Mint-2 kit, all
128 amplified cDNAs were bordered at their 5' end by the M1 sequence
129 (AAGCAGTGGTATCAACGCAGAGT) and the *Sfi*IA restriction site (GGCCATTACGGCC)
130 while, at their 3' end, they were bordered by the *Sfi*IB restriction site (GGCCGAGGCGGCC) and
131 the M1 sequence. Double stranded cDNA was purified by phenol-chloroform extraction,
132 precipitated by 2.5 volume of ethanol and 0.1 volume of sodium acetate, resuspended in ultrapure
133 water and quantified.

134 *2.2. Capture probes design and synthesis*

135 As in July 2012, all publicly available GH11 DNA coding sequences of eukaryotic origin were
136 identified by BLAST searches[28] and collected from GenBank
137 (http://www.ncbi.nlm.nih.gov/genbank/), the Joint Genome Institute database (http://jgi.doe.gov/),
138 the Broad Institute genome database (http://www.broadinstitute.org/) and CAZy
139 (http://www.cazy.org/). A set of 35 31-mers, degenerate capture probes, targeting the catalytic
140 domain of the encoded proteins (pfam no. PF00457, approx. 540 nucleotide-long; Fig. S1), was
141 designed from a collection of 342 coding DNA sequences (CDS) using the KASpOD software[11].
142 Individual probes coverage ranged from 7 to 54% of the 342 sequences, leading to a probe set
143 coverage of 90% (4 allowed mismatches).

144 The 35 oligonucleotide probes included the specific sequences (-$(X)_{31}$-) targeting cDNAs
145 encoding GH11 and adaptor sequences at each extremities for PCR amplification:
146 ATCGCACCAGCGTGT-$(X)_{31}$-CACTGCGGCTCCTCA (Table S2; Fig. S1). Biotinylated RNA
147 capture probes were prepared according to the two steps procedure of Gnirke *et al.*[29]. In the first
148 step, each single stranded DNA probe was amplified by PCR using primers complementary to the 5'

and 3' adaptors to allow double strand DNA formation. In the second step, agarose gel-purified double stranded DNA probes were converted into biotinylated RNA probes by *in vitro* transcription using the MEGAScript®T7 kit (Ambion) and biotin-dUTP (TeBu Bio). RNA probes were then mixed together in equimolar amounts.

2.3. cDNA capture

cDNA capture was carried out as described by Denonfoux *et al*.[9] and summarised in Figure S2. Briefly, 500 ng of heat denatured PCR-amplified cDNAs were hybridized to the equimolar mix of biotinylated RNA probes (500 ng) for 24 h at 65°C. Probe/cDNA hybrids were trapped by streptavidin-coated paramagnetic beads (Dynabeads® M-280 Streptavidin, Invitrogen). After different washing steps to remove unbound cDNAs, the captured cDNAs were eluted from the beads using 50 µl of 0.1 M NaOH at room temperature, neutralized with 70 µl of 1M Tris HCl pH 7.5 and purified using the Qiaquick PCR purification kit (Qiagen).

Captured cDNAs were PCR amplified using primer M1 which binds at both 5' and 3' ends of the cDNAs. PCR reactions were set up using 5 µl of eluate, 200 µM dNTPs, 400 nM primer M1, 5 µl of reaction buffer 10X (Evrogen), and 1 µl of 50X Encyclo DNA polymerase (Evrogen) in a final volume of 50 µl. After an initial denaturation at 95°C for 1 min, cDNAs were amplified for 25 cycles comprising 15 sec at 95°C, 20 sec at 66°C and 3 min at 72°C. Ten independent amplifications were conducted for each sample. PCR products of the same sample were purified on QIAquick columns (Qiagen) and pooled. A second round of hybridization and PCR amplification was performed using each of the amplified cDNA samples obtained after the first hybridization capture. Purified products originating from the same cDNA sample were pooled together and quantified by spectrophotometry (NanoDrop^{TM} 2000, Thermo Scientific). The DNA quality and size distribution of captured cDNA were assessed on an Agilent 2100 Bioanalyzer DNA 12000 chip (Agilent Technologies).

175 *2.4. Semi-quantitative PCR*

176 Enrichment in GH11 sequences at each step of the capture protocol was evaluated by semi-
177 quantitative PCR using different quantities of cDNAs and GH11-fungal specific degenerate primers
178 GH11-F (GGVAAGGGITGGAAYCNNGG) and GH11-R (TGKCGRACIGACCARTAYTG)
179 amplifying a ± 281 bp fragment (Luis et al., unpublished). PCRs were performed using either 10, 1,
180 0.1 or 0.01 ng cDNAs obtained before, after one or two cycles of hybridization capture. Twenty five
181 µl PCR reaction mixes contained 1 µl of template cDNA, 2.5 µl of 10X PCR buffer without Mg
182 (Invitrogen), 1.5 mM $MgCl_2$, 0.8 mM of each dNTP, 0.5 µM of each primer and 1 U of *Taq* DNA
183 polymerase (Invitrogen). After an initial denaturation at 94°C for 3 min, GH11 gene fragments were
184 amplified for 45 cycles comprising 45 sec at 94°C, 45 sec at 50°C and 2 min at 72°C. After a final
185 elongation at 72 °C for 10 min., 10 µl of PCR products were run in a 1.5% ethidium bromide
186 stained agarose gel.

187 *2.5. High-throughput sequencing*

188 Diversity of GH11 sequences at each step of the capture protocol was evaluated by high-
189 throughput sequencing of GH11 PCR products obtained, as described above, using primers GH11-F
190 and GH11-R. PCRs were performed using cDNAs obtained before, after one or two cycles of
191 hybridization capture. Twenty five µl PCR reaction mixes contained 10 ng of template cDNA, 2.5
192 µl of 10X PCR buffer without Mg (Invitrogen), 1.5 mM $MgCl_2$, 0.8 mM of each dNTP, 0.5 µM of
193 each primer and 1.25 U of DNA polymerase (a 24:1 mix of Invitrogen *Taq* DNA polymerase and
194 Biorad iProof polymerase). PCR cycling conditions were as described above. Five different PCR
195 reactions were prepared and run in parallel for each cDNA sample. PCR products were first
196 checked on 1.5% agarose gel before pooling together the five replicates and purification using the
197 QIAquick PCR purification kit (Qiagen). Paired-end sequencing (2x250 bp) was carried out on an
198 Illumina MiSeq sequencer (Fasteris, Switzerland).

199 Paired-end reads were assembled using PandaSeq v.2.5[30] and all sequences containing
200 unidentified nucleotide positions ("N") were filtered out. Primers and barcodes were removed using

MOTHUR v.1.30.2[31]. UCHIME[32] was used for chimera detection and sequence clusters were constructed at a 95% nucleotide sequence identity threshold. The most abundant representative sequence of each of the most abundant clusters, altogether encompassing >90% of the sequences, was translated into amino acid sequence using the ORF Finder tool of the Sequence Manipulation Suite[33] (http://www.bioinformatics.org/sms2/). Shannon diversity indices (H') were calculated after rarefying the different datasets from the same soil to the same sequencing depth (i.e. the lowest sequencing depth of the three samples of each soil). Venn diagrams were drawn using the BioVenn tool (http://www.cmbi.ru.nl/cdd/biovenn/).

2.6. Full length cDNA cloning and sequencing

Amplified cDNAs obtained after two rounds of hybridization capture were digested by *Sfi*I (Fermentas), which recognizes two distinct *Sfi*IA and *Sfi*IB sites located at the 5' and 3' ends of the cDNAs, respectively. Digested cDNAs were then ligated to the *Sfi*I-digested pDR196-SfiI-Kan yeast expression vector[34] modified to contain two *Sfi*IA and *Sfi*IB sites, downstream of the *Saccharomyces cerevisiae PMA1* promoter; thus allowing the directional cloning and potential constitutive expression of the cDNAs in yeast.

Several transformed, kanamycin-resistant *E. coli* (One Shot® TOP10 strain, Invitrogen) colonies from each sample were first randomly selected and subjected to colony PCR using the GH11-F and GH11-R primers to detect the presence of a GH11 cDNA insert. cDNA inserts from PCR-positive bacterial colonies were entirely sequenced by BIOFIDAL (Villeurbanne, France) using a PMA1 primer (CTCTCTTTTATACACACATTC) and additional internal primers when necessary.

2.7. Plasmid library construction, yeast transformation and functional screening

For each cDNA sample, a minimum of 2000 independent kanamycin-resistant transformed *E. coli* colonies were pooled together for plasmid extraction using the alkaline lysis method[35]. Aliquot samples of each plasmid library were used to transform the *S. cerevisiae* strain DSY-5 (*MATα leu2 trp1 ura3-52 his3::PGAL1-GAL4 pep4 prb1-1122*; Dualsystems Biotech) using a standard lithium

227 acetate protocol[36]. Transformed yeasts were selected on a solid yeast nitrogen base (YNB) minimal
228 medium supplemented with glucose (2%) and amino acids, but lacking uracil. YNB agar plates
229 were overlaid by a thin layer of the same medium containing 4 $mg.l^{-1}$ of AZCL-xylan (Megazyme),
230 a substrate specific for endo-xylanases. Plates were incubated at 30°C. Yeast colonies producing a
231 secreted endoxylanase were surrounded by a dark blue halo resulting from the hydrolysis of AZCL-
232 xylan.

233 For each sample, several yeast colonies positive for endoxylanase activity were picked, lysed at
234 95°C for 10 minutes in 3 µl of 20 mM NaOH and the pDR196 insert amplified by PCR using
235 primers PMA1 and ADH (GCGAATTTCTTATGATTTATG). PCR products were sequenced by
236 BIOFIDAL using the PMA1 primer.

237 *2.8. Phylogenetic analyses*

238 Sequences obtained from plasmid inserts were manually edited and corrected. Deduced amino
239 acid sequences were aligned using MUSCLE[37] to GH11 amino acid sequences obtained from public
240 databases. Maximum likelihood phylogeny analyses were generated with the PhyML 3.0 program
241 using the WAG substitution model as implemented in SeaView v. 4[38]. Phylogenetic trees were
242 drawn in MEGA v. 6[39].

243 *2.9. Sequence accessibility*

244 Sequences from plasmid inserts are available in the EBI/DDJB/GenBank databases under
245 accession Nos. LK932029-LK932091. Illumina MiSeq sequence reads have been deposited in the
246 Sequence Read Archive (SRA) of the EBI database under study No. PRJEB6672.

247

248 **3. Results**

249

250 *3.1. GH11 cDNA capture*

251 As in July 2012, we identified and collected from public databases 342 full-length eukaryotic
252 GH11 DNA coding sequences, from 113 fungal species and from two non-fungal ones. Seventy two

253 percent of these sequences were from Ascomycotina (85 species), 20% from Basidiomycotina (26
254 species) and 7% from other taxonomic groups. Prevalence of sequences from Ascomycotina is
255 likely to reflect a greater genome sequencing effort in this taxonomic group, rather than a higher
256 occurrence of the GH11 family among Ascomycotina[23]. Among the publicly available sequences,
257 those putatively full-length sequences ranged in size from 639 to 2099 bp. Occurrence of
258 carbohydrate binding motives or of C-terminal, non catalytic, extensions in the encoded
259 polypeptides accounted for most of these size variations. The 35 degenerate capture probes were
260 exclusively designed on the shared ca. 540-bp long conserved catalytic domain and were
261 susceptible to hybridize to 90% of the collected sequences.

262 SHS was performed on cDNAs synthesized from polyadenylated mRNAs extracted from four
263 different forest soils. Electrophoregrams of all cDNAs recovered after two successive rounds of
264 capture were characterised by a background smear of which emerged discrete bands ranging in size
265 from 300 to 1500 bp (Fig.1).

266 Successful enrichment in GH11 sequences along the capture protocol was demonstrated by
267 semi-quantitative PCR using GH11-specific PCR primers and different quantities of cDNA in the
268 PCR reactions (from 10 to 0.01 ng). As illustrated in Figure 2 for the Breuil beech forest (BRH
269 sample) and for the other soil samples in Figure S3, clear positive amplification of a GH11
270 fragment after two rounds of capture was always obtained using the lowest quantity of cDNA (0.01
271 ng), whereas no amplification could be observed for the same amount of cDNA prior to SHS.

272 *3.2. Cloning, sequencing and heterologous expression of captured cDNA*

273 Captured cDNAs in the range of 700-1500 bp were cloned into the pDR196 *E. coli* / *S.*
274 *cerevisiae* shuttle expression vector to constitute four soil-specific GH11-enriched plasmid libraries
275 (Table 1). Forty recombinant colonies per library were randomly screened by PCR using GH11-
276 specific primers to evaluate the percentage of GH11-containing recombinant plasmids. Efficient
277 enrichment occurred for all libraries with 80 to more than 90% of positive clones (Table 1). Among
278 the 55 fully sequenced plasmid inserts from PCR positive colonies, all but two indeed corresponded

279 to GH11 sequences (Table 1). Seventy two percent of the sequences encoded putatively full-length
280 GH11 polypeptides based on alignment length to known GH11 polypeptides and the presence of in-
281 frame putative start and stop codons. Out of them, 15% were characterised by the presence of a
282 family 1 carbohydrate binding domain (CBM1) in C-terminal position.

283 Functional screening using *S. cerevisiae* was conducted on the four GH11-enriched plasmid
284 libraries by plating the recombinant yeasts onto a medium supplemented with an endoxylanase-
285 specific colour reagent (AZCL-xylan). Depending on the library, between 1.5 (sample PUE) and
286 25% (sample BRH) of the transformed yeast colonies developed a dark blue halo demonstrating
287 secretion of a functional endo-xylanase (Fig. S4). All eleven sequenced plasmid inserts from these
288 xylanase-positive yeast colonies encoded GH11 proteins; five of them had already been identified
289 among sequences obtained from bacterial colonies and four had a C-terminal CBM1 domain.

290 *3.3. Selectivity of the SHS GH11 capture*

291 To evaluate the diversity of GH11 sequences at each step of the capture protocol we performed
292 a high-throughput Illumina MiSeq sequencing of GH11 amplicons obtained from all four cDNA
293 samples, prior (H0) and after one (H1) or two (H2) cycles of SHS capture. Paired-end sequence
294 reads were assembled to reconstitute the ca. 281 bp-long amplicons. Altogether, the total dataset
295 contained 334,161 full-length amplicon sequences which were clustered at a 95% nucleotide
296 sequence identity threshold to produce a total number of 1458 clusters, of which 1001 (69%) were
297 singletons (data summarized in Table 2 for each sample). Each of the 12 sequence datasets (4
298 cDNA samples x the 3 steps of the SHS) was characterised by few dominant clusters encompassing
299 most of the sequences and a large number of clusters each containing a few, or even a single,
300 sequences (illustrated in Fig. 3A for the PUE sample). None of the sequences obtained were
301 identical to sequences deposited in databases. Only 17 of the sequence clusters of which 14
302 exclusively from the BEW site, were more than 90% identical (maximum value of 97.5%) at the
303 nucleotide level over their entire length to GH11 genes from either the Basidiomycota *Tulasnella*
304 *calospora* or the Ascomycota *Nectria haematococca* and *Pyrenophora teres*.

Figure 3 also showed that the most abundant sequence clusters obtained after one (H1) and two (H2) cycles of capture did not, for a majority of them, correspond to the most abundant clusters present before capture (H0). Venn diagrams drawn using only these most prominent sequence clusters, encompassing altogether 90-93% of a sample sequences, showed that it existed a larger overlap between the post-capture samples H1 and H2 than between the pre-capture sample H0 and H1 or H2 (Figure 3B). This trend was observed to some extent for samples BEW, BRE and PUE, but not for the BRH one which differed from the others by the dominance of only three clusters in the H0 cDNA pool which encompassed 90% of the sequenced reads (Fig. S5). Despite these apparent differences in sequence distribution between the pre-capture H0 and the post capture H1 and H2 samples, sequence diversity indices, such as the Shannon index, did not differ between the pre- and post-capture sequence pools (Table 2, with the exception of the BRH sample). Between 2.7% (BRE and BEW) and 15% (PUE and BRH) of the sequence clusters were shared between two sites. Eight sequence clusters were identified in all 4 studied sites.

To address the phylogenetic diversity of the captured sequences we first produced an amino acid sequence alignment of 62 known GH11 proteins representative of the phylogenetic diversity of this gene family. To this alignment we added the GH11 sequences obtained by the random sequencing of plasmid inserts, the sequences producing a functional enzyme in yeast and the sequences representative of the most abundant Illumina sequence clusters before (H0) or after (H1 and H2) SHS capture. The GH11 family is a highly diversified and fast evolving gene family and phylogenies based either on full-length protein sequence alignments or on partial alignments, as in the present case, clearly do not reflect the species phylogenies and comprise very few well supported internal branches (Figure 4). Phylogenetic trees obtained for sequences from the four studied soils (Fig. 4 and Fig. S6) all clearly showed that the captured sequences were distributed over the entire reference tree.

4. Discussion

The results obtained clearly demonstrate that solution hybridization selection (SHS) represents a powerful strategy to select full-length cDNAs, representative of a specific gene family, originally diluted in a highly complex metatranscriptomic sequence pool. This protocol was successfully implemented on four different forest soil RNA samples. Based on previous estimates of the frequency of GH11 sequences among eukaryotic cDNA for two of the soils used in the present study (BRE and BRH)[22], two successive cycles of SHS have the potential to enrich specific cDNA sequences by a factor of at least 10^4. As suggested by the results of the semi-quantitative PCR, in some cases (e.g. PUE sample, Fig. S3), one cycle of capture may be sufficient to get a maximum level of enrichment, while in other cases two cycles seem required (e.g. BRH sample, Fig. 2):

Sequence analysis of PCR fragments amplified from pre- or post-capture cDNAs demonstrated that capture succeeded in selecting both a large number and phylogenetically diverse representatives of the selected gene family. Furthermore, none of the captured sequences appeared to be identical to already known ones which were originally used for probe design. Capture, could however preferentially select sequences which were not necessarily among the most abundant in the original cDNA pool. Such results should be evaluated in the future by quantitative PCR assays. Despite explorative probe design strategy, publicly available homologous sequences at the start of the study greatly influence the capture selectivity. Probe sets utilized to capture a given biomarker should therefore be upgraded regularly taking into account newly deposited sequences. Thanks to the ever increasing number of published fungal genomes, representative of the phylogenetic diversity of this taxonomic group, explorative probe design strategies could be carried out to unravel the metabolic capacities of these microorganisms within different ecosystems. Besides GH11 sequences, SHS capture can be implemented for any other gene family of interest allowing a comprehensive taxonomic or functional description of the studied microbial community.

As mentioned in the introduction, sequence capture presents the advantage over PCR to give access to the full-length gene sequence, including facultative modules, not always associated to the

studied catalytic domain. This was indeed the case for the GH11, for which we estimated that 72% of the captured sequences were full length and that 15% of them processed a C-terminal, fungal specific, CBM1 module (see the CAZy database, http://www.cazy.org). A discrepancy however existed between the estimated fraction of full-length captured GH11 cDNA and the systematically lower fraction of cDNAs which produced a functional enzyme upon expression in *S. cerevisiae*. Absence of expression in yeast can be attributed to a number of independent factors ranging from bias in codon usage, non-recognition by *S. cerevisiae* of the protein signal peptide necessary for correct secretion, protein misfolding or hyperglycosylation. Some of these problems could be addressed by using expression plasmids including a yeast signal peptide downstream of the cloning site and/or by using a different yeast species for protein production.

Sequencing of PCR fragments amplified from captured cDNAs also indicate that altogether the four captured cDNA samples obtained in this single study encompass a greater number of novel and different GH11 sequences than have been deposited and are available in public databases over several decades. This observation should promote the use of cDNA sequence capture (i) as a complementary approach to PCR to explore and quantify the extent of eukaryotic functional diversity in complex environments, but also (ii) as a powerful tool in environmental biotechnology to efficiently screen for enzyme variants with novel biochemical properties.

Acknowledgements

We would like to thank Richard Joffre, Jacques Ranger and Alberto Orgiazzi for soil sampling at the Puéchabon, Breuil and Berchidda sites respectively. Audrey Dubost and Stefano Ghignone contributed to bioinformatics analyses and Jérémie Denonfoux for his valuable help with the gene capture. We acknowledge the JGI of the US Department of Energy and the Broad Institute for making available genome data prior to their publication. CB was supported by the University of Torino and the région Rhône-Alpes (CMIRA program); CR received a graduate grant from the Ministère de l'Enseignement Supérieur et de la Recherche; NP was funded by the Direction

Générale de l'Armement and EP by the Agence Nationale pour la Recherche. Work was financed by the CNRS-INSU ECCO Microbien program, the INRA métaprogramme M2E (project Metascreen), project ANR 09-GENM-033-001 (Eumetasol); EU-project 'EcoFINDERS' No. 264465 and local funding by the University of Torino (ex-60%).

References

1. Tringe, S.G., Rubin, E.M. 2005, Metagenomics: DNA sequencing of environmental samples, *Nat. Rev. Genet.,* 6(11), 805-814.

2. Simon, C., Daniel, R. 2011, Metagenomic analyses: past and future trends, *Appl. Environ. Microbiol*., 77(4), 1153-1161.

3. Howe, A.C., Jansson, J.K., Malfatti, S.A., Tringe, S.G., Tiedje, J.M., Brown, C.T. 2014, Tackling soil diversity with the assembly of large, complex metagenomes, *Proc. Natl. Acad. Sci. U S A*, 111(13), 4904-4909.

4. Voříšková, J., Baldrian, P. 2013, Fungal community on decomposing leaf litter undergoes rapid successional changes, *ISME J.*, 7(3), 477-486.

5. Luis, P., Kellner, H., Zimdars, B., Langer, U., Martin, F., Buscot, F. 2005, Patchiness and spatial distribution of laccase genes of ectomycorrhizal, saprotrophic, and unknown basidiomycetes in the upper horizons of a mixed forest cambisol, *Microb. Ecol.*, 50(4), 570-579.

6. Kellner, H., Luis, P., Pecyna, M.J., et al. 2014, Widespread occurrence of expressed fungal secretory peroxidases in forest soils, PLoS ONE, 9(4):e95557.

7. Kellner, H., Zak, D.R., Vandenbol, M. 2010, Fungi unearthed: transcripts encoding lignocellulolytic and chitinolytic enzymes in forest soil, PLoS ONE, 5(6):e10971. doi: 10.1371/journal.pone.0010971. Erratum in: PLoS One,5(9).

8. Hong, S., Bunge, J., Leslin, C., Jeon, S., Epstein, S.S. 2009, Polymerase chain reaction primers miss half of rRNA microbial diversity, *ISME J.*, 3(12), 1365-1373.

9. Denonfoux, J., Parisot, N., Dugat-Bony, E., et al. 2013, Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration DNA, *DNA Res.,* 20(2), 185-196.

10. Dugat-Bony, E., Missaoui, M., Peyretaillade, E., et al. 2011, HiSpOD: probe design for functional DNA microarrays, *Bioinformatics*, 27(5), 641–648.

11.Parisot, N., Denonfoux, J., Dugat-Bony, E., Peyret, P., Peyretaillade, E. 2012, KASpOD-a web service for highly specific and explorative oligonucleotide design, *Bioinformatics,* 28(23), 3161–3162.

12. Bailly, J., Fraissinet-Tachet, L., Verner, M.C., et al. 2007, Soil eukaryotic functional diversity, a metatranscriptomic approach, *ISME J.*, 1(7), 632–642.

13. Kellner, H., Luis, P., Portetelle, D., Vandenbol, M. 2011, Screening of a soil metatranscriptomic library by functional complementation of *Saccharomyces cerevisiae* mutants, *Microbiol. Res*., 166(5), 360-368.

14. Damon, C., Vallon, L., Zimmermann, S., et al. 2011, A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes, *ISME J.*, 5(12), 1871-1880.

15. Lehembre, F., Doillon, D., David, E., et al. 2013, Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes*, Environ. Microbiol*., 15(10), 2829–2840.

16. Takasaki, K., Miura, T., Kanno, M., et al. 2013, Discovery of glycoside hydrolase enzymes in an avicel-adapted forest soil fungal community by a metatranscriptomic approach, PLoS ONE, 8(2):e55485.

17. Bates, S.T., Clemente, J.C., Flores, G.E., et al. 2013, Global biogeography of highly diverse protistan communities in soil, *ISME J.,* 7(3), 652-659.

18. Taylor, D. L., Hollingsworth, T. N., McFarland, J.W., Lennon, N. J., Nusbaum, C., Ruess, R.W. 2014, A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning, *Ecological Monograph,* 84(1), 3–20.

19. Schneider, T., Keiblinger, K.M., Schmid, E., et al. 2012, Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions, *ISME J*., 6(9), 1749–1762.

20. Stursová, M., Zifčáková, L., Leigh, M.B., Burgess, R., Baldrian, P. 2012, Cellulose utilization in forest litter and soil: identification of bacterial and fungal decomposers, *FEMS Microbiol. Ecol.*, 80(3), 735-746.

21. Demain, A.L., Dana, C.A. 2007, The business of biotechnology, *Industrial Biotechnology*, 3, 269-283.

22. Damon, C., Lehembre, F., Oger-Desfeux, C., et al. 2012, Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils, *PloS ONE*, 7(1), e28967.

23. Grigoriev IV, Nikitin, R., Haridas, S., et al. 2014, MycoCosm portal: gearing up for 1000 fungal genomes, *Nucleic Acids Res*., 42(Database issue), D699-704.

24. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B. 2014, The Carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic Acids Res.* 42, D490–495.

25. Paës, G., Berrin J. G., Beaugrand, J. 2012, GH11 xylanases: Structure/function/properties relationships and applications, *Biotechnol. Adv.*, 30 (3), 564-592.

449 26. Kuramae, E.E., Hillekens, R.H., de Hollander, M., van der Heijden, M.G., van den Berg, M.,
450 van Straalen, N.M., Kowalchuk, G.A. 2013, Structural and functional variation in soil fungal
451 communities associated with litter bags containing maize leaf, *FEMS Microbiol. Ecol.* 84(3), 519-
452 531.
453 27. Damon, C., Barroso, G., Férandon, C., Ranger, J., Fraissinet-Tachet, L., Marmeisse, R. 2010,
454 Performance of the COX1 gene as a marker for the study of metabolically active Pezizomycotina
455 and Agaricomycetes fungal communities from the analysis of soil RNA, *FEMS Microbiol. Ecol.*,
456 74(3), 693–705.
457 28. Altschul, S.F., Gish, W., Miller, W., Myers, E. W., Lipman, D.J. 1990, Basic local alignment
458 search tool, *J. Mol. Biol.*, 215(3), 403-410.
459 29. Gnirke, A., Melnikov, A., Maguire, J.et al. 2009, Solution hybrid selection with ultra-long
460 oligonucleotides for massively parallel targeted sequencing, *Nat. Biotechnol*., 27(2), 182–189.
461 30. Masella, A. P., Bartram, A. K., Truszkowski, J.M., Brown, D. G., Neufeld, J. D. 2012,
462 PANDAseq: paired-end assembler for illumina sequences, *BMC Bioinformatics*, 13:31.
463 31. Schloss, P. D., Westcott, S.L., Ryabin, T., et al. 2009, Introducing mothur: open-source,
464 platform-independent, community-supported software for describing and comparing microbial
465 communities, *Appl. Environ. Microbiol*. 75(23), 7537–7541.
466 32. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R. 2011, UCHIME improves
467 sensitivity and speed of chimera detection, *Bioinformatics* 27(16), 2194–2200.
468 33. Stothard, P. 2000, The sequence manipulation suite: JavaScript programs for analyzing and
469 formatting protein and DNA sequences. *Biotechniques* 28(6),1102–1104.
470 34. Rentsch, D., Laloi, M., Rouhara, I., Schmelzer, E., Delrot, S., Frommer, W. B. 1995, NTR1
471 encodes a high affinity oligopeptide transporter in Arabidopsis, *FEBS Lett.*, 370(3), 264–268.
472 35. Sambrook, J., Russell, D.W. 2001, Molecular Cloning: a Laboratory Manual. 3rd Ed., Cold
473 Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
474 36. Rose, M., Winston, F., Hieter, P. 1990, Methods in Yeast Genetics: a Laboratory Course
475 Manual. Cold Spring Harbor Laboratory Press , Cold Spring Harbor, New York.
476 37. Edgar, R. C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high
477 throughput, *Nucleic Acids Res.,* 32(5), 1792-1797.
478 38. Gouy, M., Guindon, S., Gascuel, O. 2010, SeaView version 4: A multiplatform graphical user
479 interface for sequence alignment and phylogenetic tree building, *Mol. Biol. Evol.,* 27(2), 221–224.
480 39. Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. 2013, MEGA6: Molecular
481 Evolutionary Genetics Analysis Version 6.0, *Mol. Biol. Evol*., 30(12), 2725-2729.
482

Tables

Table 1. Cloning and characterization of captured GH11 cDNAs

Samples	**PUE**	**BRH**	**BRE**	**BEW**
No. of captured cDNAs cloned in *E. coli*	6770	2020	5720	5880
No. of *E. coli* colonies screened by PCR	40	40	40	40
Positive amplification of a GH11 fragment (%)	37(92.5)	33(82.5)	35(87.5)	36(90)
No. of inserts sequenced	12	13	16	14
No. of GH11 inserts (%)	11 (92)	12 (92)	16 (100)	14 (100)
No. of putative full length GH11(%)	9 (82)	9 (75)	11 (69)	9 (64)
% of endoxylanase-positive yeast colonies	1.5	25	12	6

Table 2 Summary statistics from Illumina MiSeq sequencing of GH11 PCR fragments amplified, for each four cDNA samples, before (H0) or after one (H1) or two (H2) hybridization capture.

Sample	Total no. of sequences	Total no. of clusters[1] (95%)	No. of clusters encompassing ≥90% of the sequences	Shannon diversity index (H')[2]	No. of shared clusters between H0-H1-H2[2]
PUE_H0	12960	298	52 (17%)	3.819	
PUE_H1	24565	227	51 (22%)	4.015	70 (11%)
PUE_H2	25053	291	46 (16%)	3.912	
BRE_H0	13538	87	9 (10%)	2.254	
BRE_H1	42000	140	5 (4%)	1.651	11 (5%)
BRE_H2	46626	112	6 (5%)	1.73	
BRH_H0	2765	26	3 (12%)	1.061	
BRH_H1	28366	51	3 (6%)	1.234	5 (4%)
BRH_H2	17322	159	18 (11%)	2.135	
BEW_H0	41799	214	15 (7%)	2.761	
BEW_H1	42308	249	10 (4%)	2.496	38 (6%)
BEW_H2	36859	205	6 (3%)	2.196	

[1] Including singletons

[2] Shannon diversity indices and shared clusters were calculated after rarefying the different datasets from the same soil to the same sequencing depth

494 **Figures legends**

495 **Figure 1.** Electrophoretic separation of cDNAs obtained following two consecutive solution
496 hybridization selection. Captured cDNAs from the 4 soil samples PUE, BRH, BRE and BEW
497 were run on an Agilent DNA 12000 microfluidic chip. Each band could encompass one or
498 several unique but abundant GH11 cDNAs.

499 **Figure 2.** Semi-quantitative PCR amplification of a 281 bp GH11 fragment using different
500 quantities (from 10 to 0.01 ng) of BRH cDNA obtained before (H0) and after one (H1) or two
501 (H2) cycles of hybridization. Before capture PCR products could only be obtained using 10
502 ng of input cDNA. Amplifications of the PUE, BRE and BEW samples are illustrated in
503 Figure S3.

504 **Figure 3.** Selectivity of the Solution Hybrid Selection (SHS) capture. (A) Rank-abundance
505 distribution of the most abundant GH11 nucleotide sequence clusters identified before (H0),
506 or after one (H1) or two (H2) cycles of hybridization on the PUE cDNAs. Only clusters
507 encompassing 80% of the sequences in the H0, H1 or H2 samples are shown. "C" or "Y"
508 letters above bars indicate sequences obtained by random sequencing of plasmid inserts or
509 which could be functionally expressed in yeast, respectively. (B) Venn diagram showing the
510 number of unique or shared GH11 sequence clusters, before (H0), or after one (H1) or two
511 (H2) cycles of hybridization on the PUE cDNAs. As in (A), only the most abundant clusters,
512 encompassing 90% of the sequences, were used for the calculation. GH11 PCR sequences
513 were clustered using a nucleotide sequence identity threshold of 95%. Similar Venn diagrams
514 for the BRH, BRE and BEW samples are illustrated in Figure S5.

515 **Figure 4.** Phylogenetic diversity of the GH11 partial amino acid sequences obtained from
516 PUE cDNA samples. 0, 1 and 2, translated PCR sequences obtained before or after one or two
517 cycles of hybridization. PUE sequences are scattered over the entire tree which includes
518 representative reference sequences from Ascomycota and Basidiomycota. c, sequences

519 obtained from *E. coli* clones; y, sequences functionally expressed in yeast clones. PhyML tree
520 calculation was based on an alignment of ca. 80 amino acid long GH11 partial sequences.
521 Thicker internal black branches indicate bootstrap value ≥60% (1000 replications). Full
522 species names and accession numbers of the reference sequences are given in Fig. S6A.
523 Similar trees drawn using the sequences from sites BRE, BRH and BEW are illustrated in Fig.
524 S6 B, C and D, respectively.
525

526 **Supplementary figures legends**

527 **Figure S1.** Position of the 35 capture probes along the GH11 catalytic domain.

528 **Figure S2.** Overview of the SHS capture method implemented in the present study. The first
529 two steps were performed twice consecutively.

530 **Figure S3.** Semi-quantitative PCR amplification of a 281 bp GH11 fragment using different
531 quantities (from 10 to 0.01 ng) of PUE, BRE or BEW cDNAs obtained before (H0) and after
532 one (H1) or two (H2) cycles of hybridization.

533 **Figure S4.** Several of the yeast colonies transformed with the plasmid library prepared from
534 captured BRH cDNAs express a functional secreted endo-xylanase. Following
535 transformation, DSY-5 yeast cells were plated on a selective medium without uracil and
536 containing AZCL-xylan, an endoxylanase-specific substrate, whose degradation leads to the
537 release of a dark blue dye.

538 **Figure S5.** Selectivity of the Solution Hybrid Selection (SHS) capture. Venn diagram
539 showing the number of unique or shared GH11 sequence clusters, before (H0), or after one
540 (H1) or two (H2) SHS capture on the BRH, BRE and BEW cDNAs. For each of the three soil
541 cDNA samples, only the most abundant sequence clusters, encompassing ≥90% of the
542 sequences in the H0, H1 or H2 samples, were used for the calculation. GH11 PCR sequences
543 were clustered using a nucleotide sequence identity threshold of 95%.

544 **Figure S6.** Phylogenetic diversity of the GH11 partial amino acid sequences obtained from
545 (A) the PUE, (B) the BRE, (C) the BRH and (D) the BEW cDNA samples (green and black
546 labels). 0, 1 and 2, translated PCR sequences obtained before or after one or two cycles of
547 hybridization. Environmental cDNA sequences are scattered over the entire tree which
548 includes representative reference sequences from Ascomycota (blue lines) and Basidiomycota
549 (red lines). c, sequences obtained from *E. coli* clones; y: sequences functionally expressed in
550 yeast clones. PhyML tree calculation was based on an alignment of ca. 80 amino acid long

551 GH11 partial sequences. Thicker black branches indicate bootstrap value ≥ 60% (1000
552 replications).

Figure 1

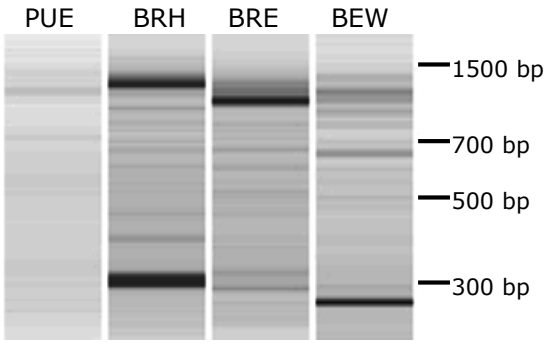


Figure 2

Figure 4

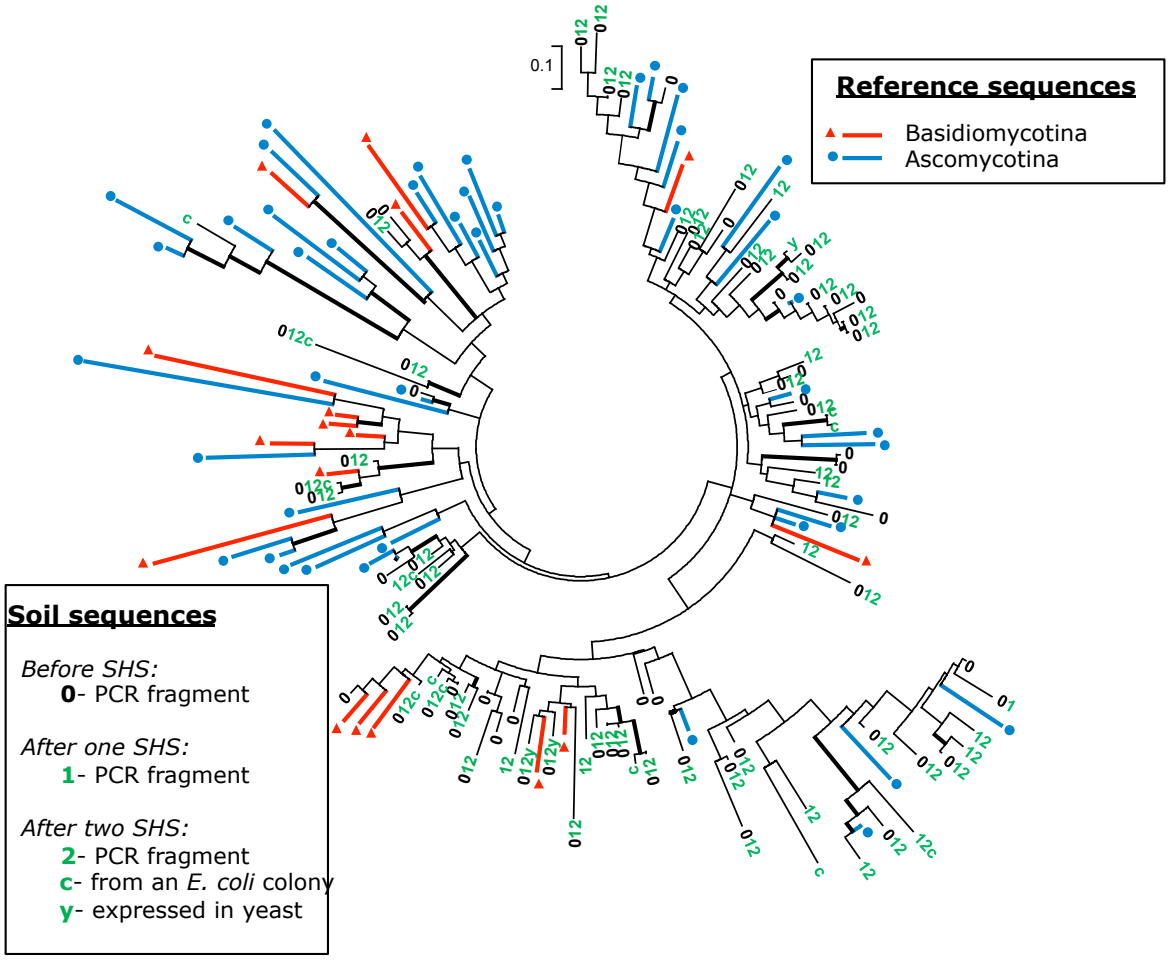

0.1
Reference sequences
Basidiomycotina
Ascomycotina
Soil sequences
Before SHS:
0- PCR fragment
After one SHS:
1- PCR fragment
After two SHS:
2- PCR fragment
c- from an E. coli colony
y- expressed in yeast

Figure S1

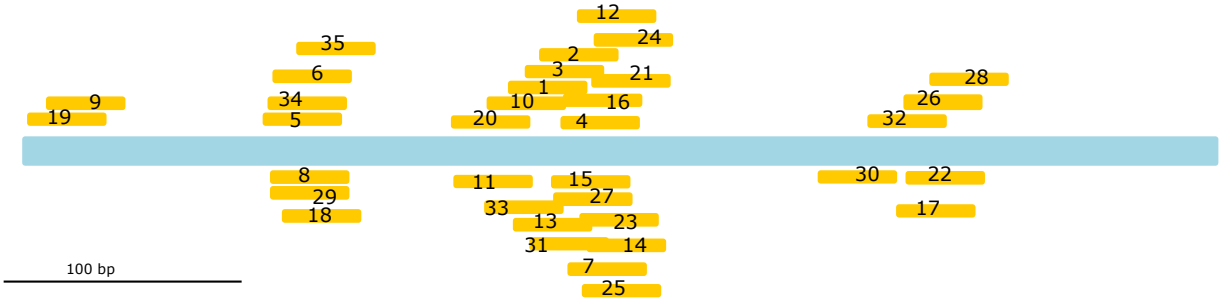

12
24
35
2
6
3
21
28
1
9
34
10
16
26
19
5
20
4
32
8
11
15
30
22
29
27
33
17
18
13
23
31
14
7
25
100 bp

Figure S2

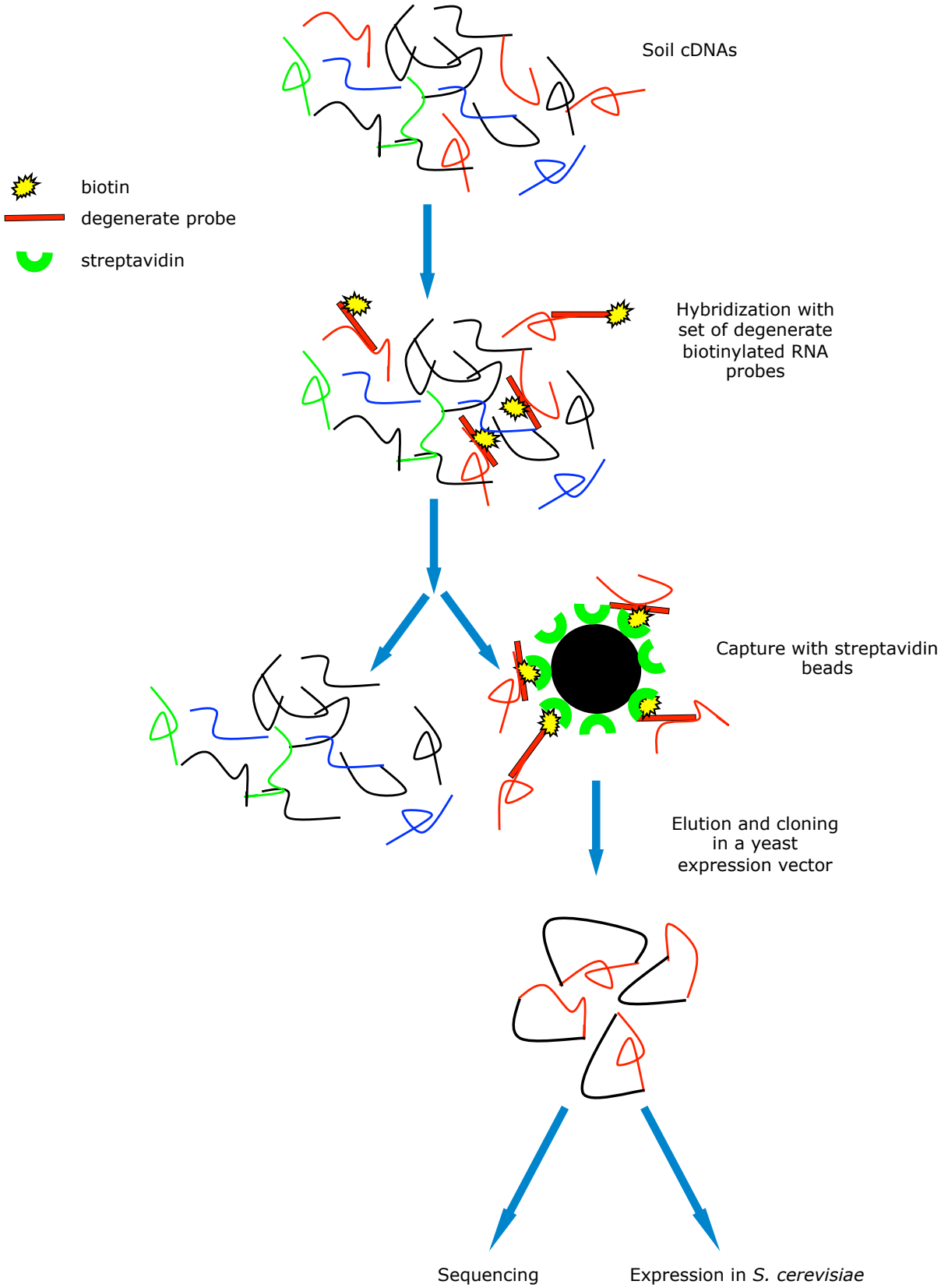

biotin
degenerate probe
streptavidin
Soil cDNAs
Hybridization with set of degenerate biotinylated RNA probes
Capture with streptavidin beads
Elution and cloning in a yeast expression vector
Sequencing
Expression in S. cerevisiae

Figure S3.

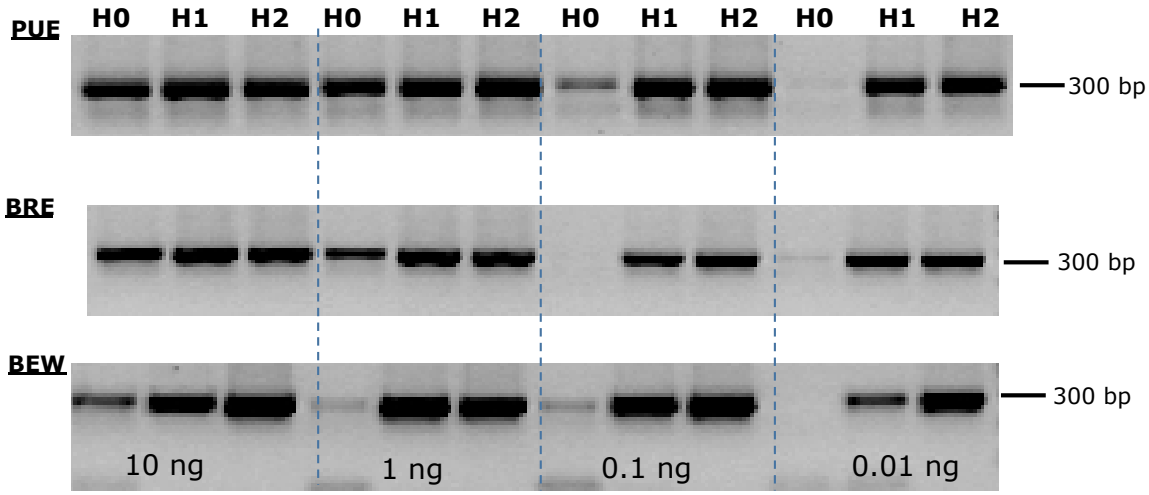

H0
H1
H2
H0
H1
H2
H0
H1
H2
H0
H1
H2
PUF
300 bp
BRE
300 bp
BEW
300 bp
10 ng
1 ng
0.1 ng
0.01 ng

Figure S4

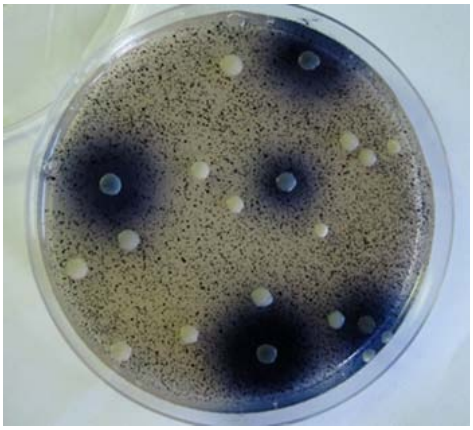
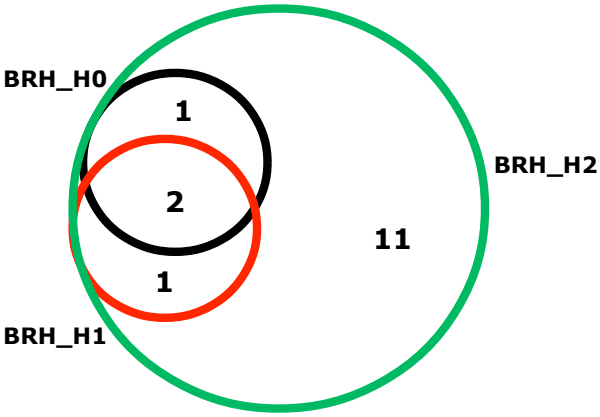
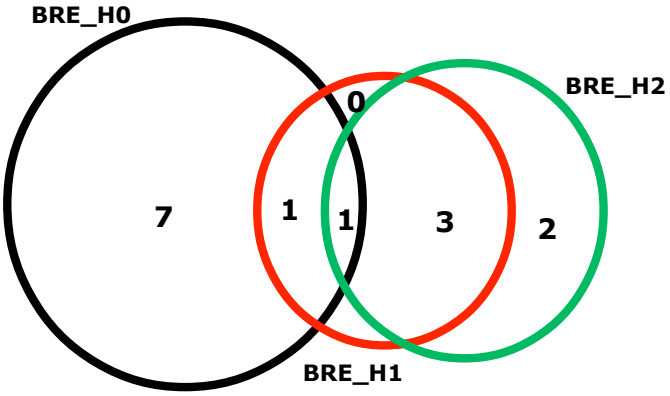


Figure S5


BRH_H0
1
2
1
11
BRH_H2
BRH_H1


BRE_H0
7
0
1
1
3
2
BRE_H2
BRE_H1

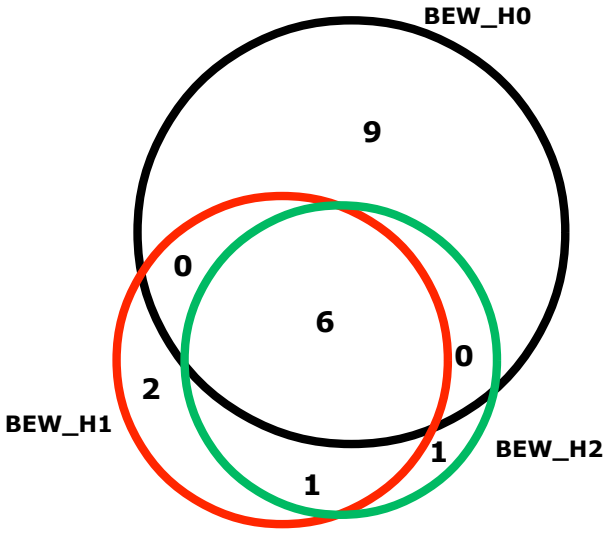
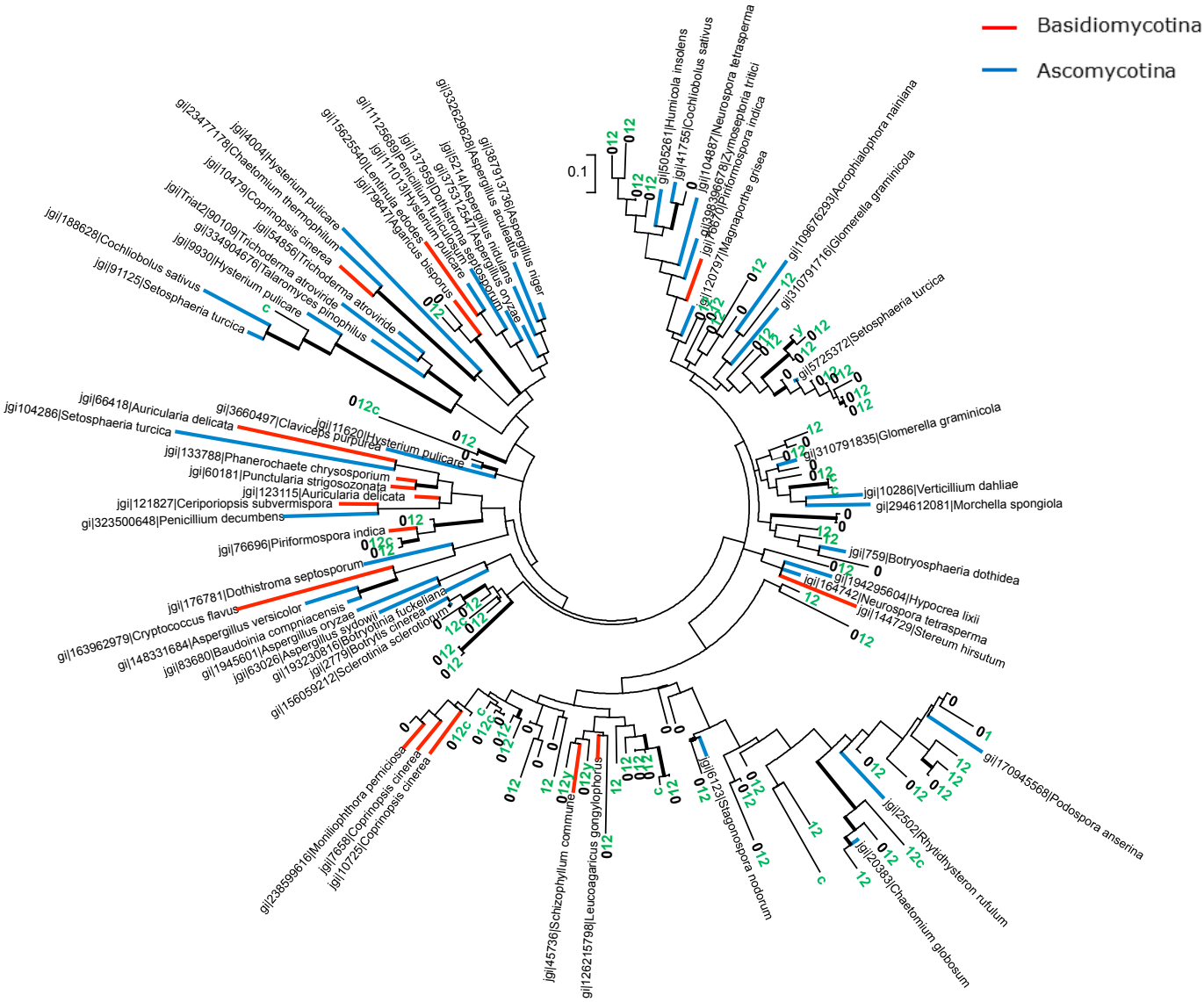
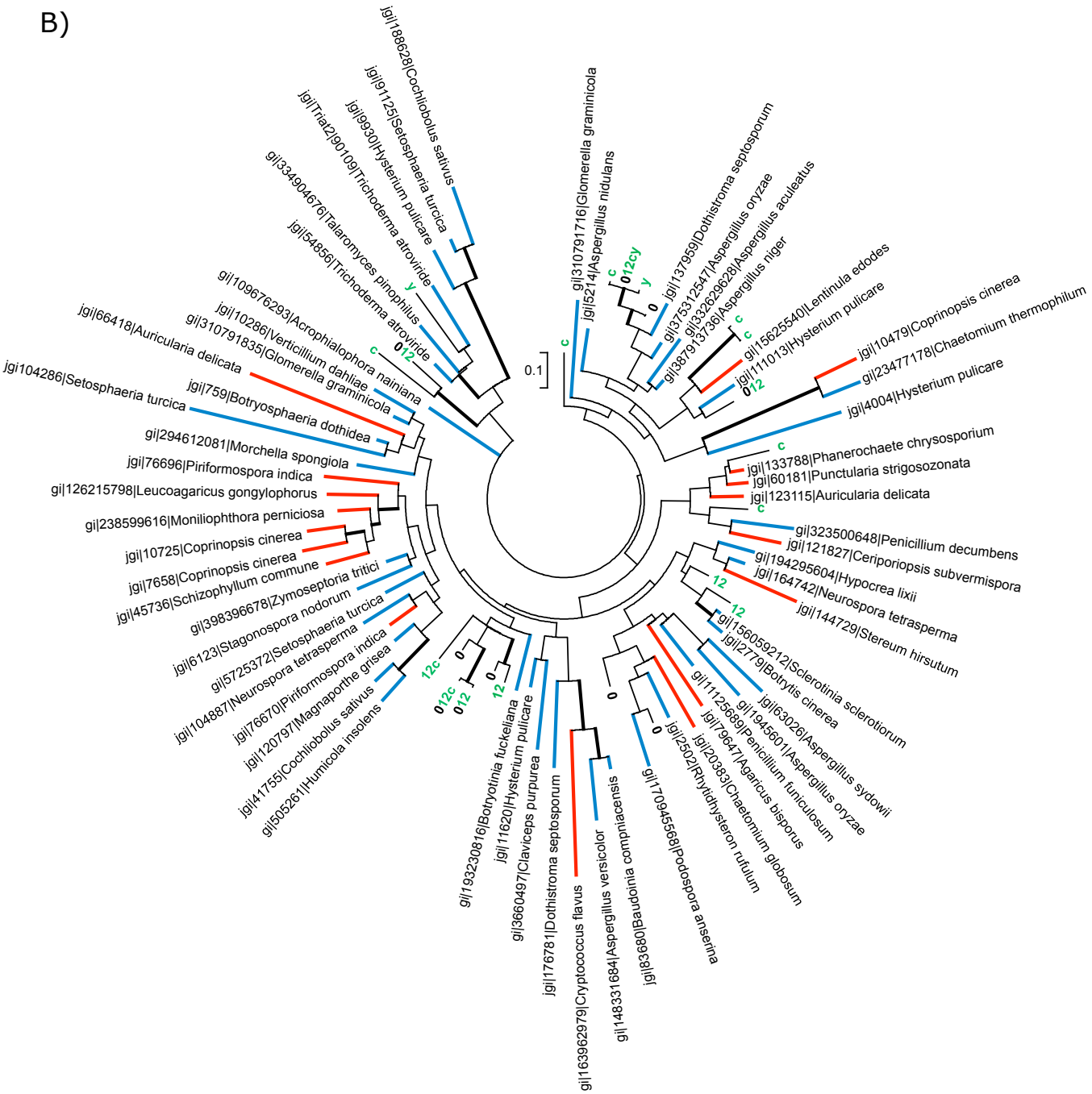

BEW_H0
9
0
6
0
2
1
1
BEW_H1
BEW_H2

Figure S6

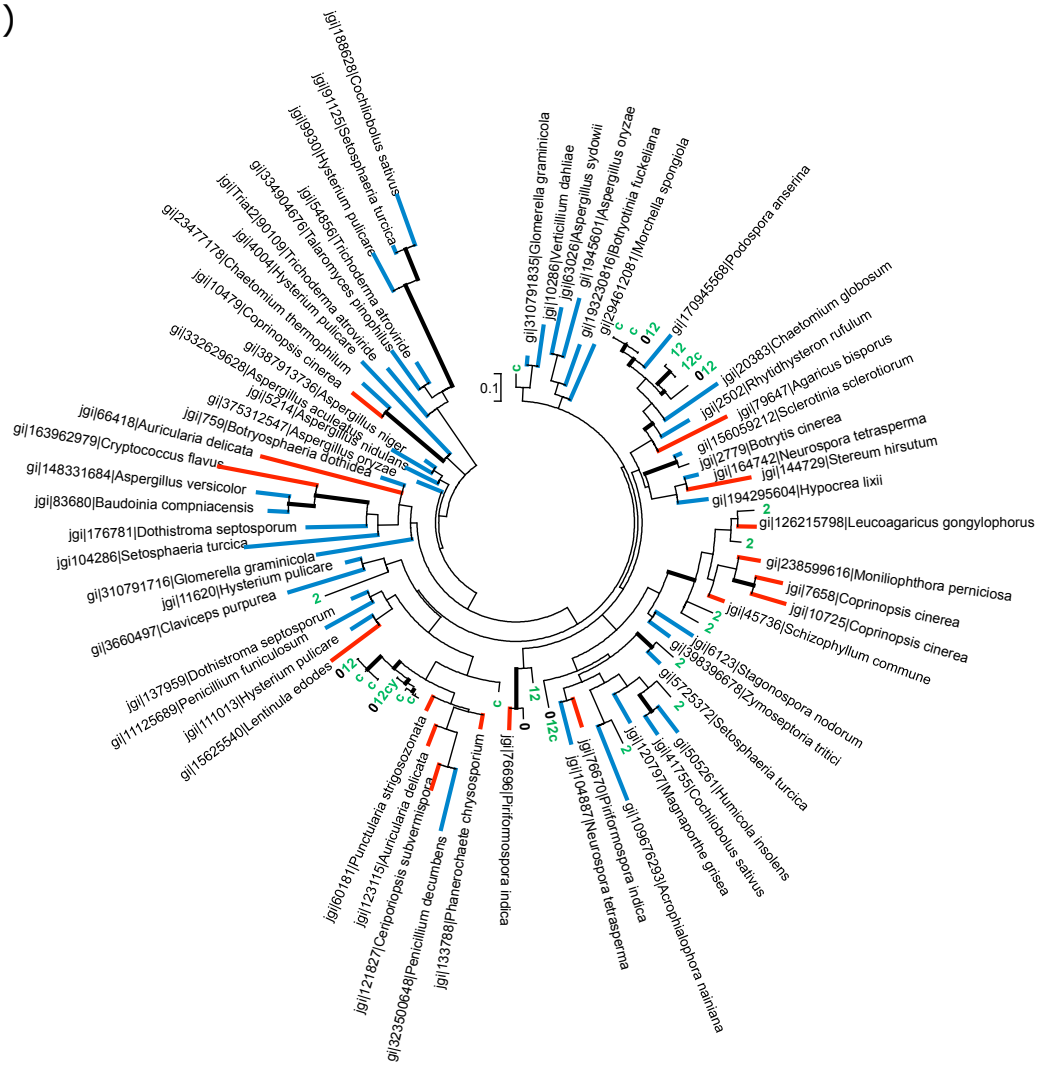
A)


Basidiomycotina
Ascomycotina
0.1
gi|50261|Humicola insolens
gi|41755|Cochliobolus sativus
gi|1104887|Neurospora tetrasperma
gi|39396678|Zymoseptoria tritici
gi|68870|Pirifomospora indica
gi|120797|Magnaporthe grisea
gi|10967629|Acrophialophora nainiana
gi|310791716|Glomerella graminicola
gi|5725372|Setosphaeria turcica
gi|310791835|Glomerella graminicola
jgi|10286|Verticillium dahliae
gi|294612081|Morchella spongiola
jgi|759|Botryosphaeria dothidea
gi|194296604|Hypocrea lixii
jgi|164742|Neurospora tetrasperma
jgi|144729|Stereum hirsutum
gi|170945568|Podospora anserina
jgi|2502|Rhytidhysteron rufulum
jgi|20383|Chaetomium globosum
jgi|6123|Stagonospora nodorum
gi|126215798|Leucoagaricus gongylophorus
jgi|45736|Schizophyllum commune
gi|238599616|Moniliophthora perniciosa
jgi|7658|Coprinopsis cinerea
jgi|10725|Coprinopsis cinerea
gi|156059212|Sclerotinia sclerotiorum
jgi|2779|Botrytis cinerea
gi|193230816|Botrytis fuckeliana
jgi|63026|Aspergillus oryzae
gi|1945601|Aspergillus sydowii
jgi|83680|Baudoinia compniacensis
gi|148331684|Aspergillus versicolor
gi|163962979|Cryptococcus flavus
jgi|176781|Dothistroma septosporum
jgi|76696|Piriformospora indica
gi|323500648|Penicillium decumbens
jgi|121827|Ceriporiopsis subvermispora
jgi|123115|Auricularia delicata
jgi|60181|Punctularia strigosozonata
jgi|133788|Phanerochaete chrysosporium
jgi|104286|Setosphaeria turcica
jgi|66418|Auricularia delicata
gi|3660497|Claviceps purpurea
jgi|11620|Hysterium pulicare
jgi|91125|Setosphaeria turcica
gi|188628|Cochliobolus sativus
jgi|9930|Hysterium pulicare
gi|334904676|Talaromyces pinophilus
jgi|Triat2|90109|Trichoderma atroviride
gi|54856|Trichoderma atroviride
jgi|10479|Coprinopsis cinerea
jgi|4004|Hysterium pulicare
gi|23477178|Chaetomium thermophilum
gi|156255540|Lentinula edodes
gi|111013|Hysterium pulicare
jgi|79641|Agaricus bisporus
jgi|137959|Penicillium funiculosum
gi|37531247|Aspergillus oryzae
gi|5214|Aspergillus aculeatus
gi|332629628|Aspergillus niduluns
gi|38791736|Aspergillus niger

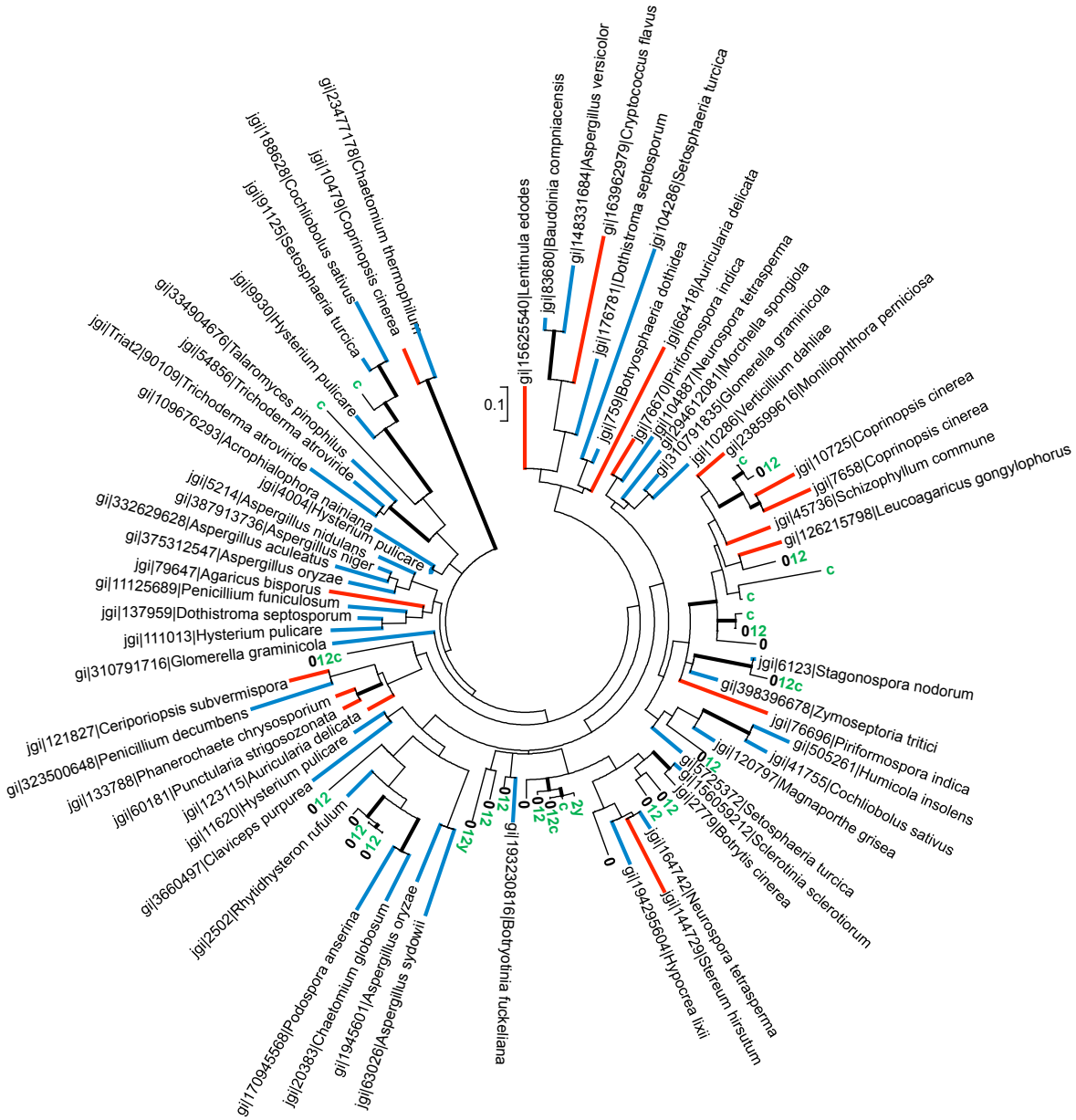
B)


jgi|188628|Cochliobolus sativus
jgi|91125|Setosphaeria turcica
jgi|9930|Hysterium pulicare
jgi|Triat2|90109|Trichoderma atroviride
gi|334904676|Talaromyces pinophilus
jgi|54856|Trichoderma atroviride
gi|109676293|Acrophialophora nainiana
jgi|10286|Verticillium dahliae
gi|310791835|Glomerella graminicola
jgi|66418|Auricularia delicata
jgi104286|Setosphaeria turcica
jgi|759|Botryosphaeria dothidea
gi|294612081|Morchella spongiola
jgi|76696|Piriformospora indica
gi|126215798|Leucoagaricus gongylophorus
gi|238599616|Moniliophthora perniciosa
jgi|10725|Coprinopsis cinerea
jgi|7658|Coprinopsis cinerea
jgi|45736|Schizophyllum commune
gi|398396678|Zymoseptoria tritici
jgi|6123|Stagonospora nodorum
gi|5725372|Setosphaeria turcica
jgi|104887|Neurospora tetrasperma
jgi|76670|Piriformospora indica
jgi|120797|Magnaporthe grisea
jgi|41755|Cochliobolus sativus
gi|505261|Humicola insolens
gi|310791716|Glomerella graminicola
jgi|5214|Aspergillus nidulans
jgi|137959|Dothistroma septosporum
gi|375312547|Aspergillus oryzae
gi|332629628|Aspergillus aculeatus
gi|387913736|Aspergillus niger
gi|15625540|Lentinula edodes
jgi|111013|Hysterium pulicare
jgi|10479|Coprinopsis cinerea
gi|23477178|Chaetomium thermophilum
jgi|4004|Hysterium pulicare
jgi|133788|Phanerochaete chrysosporium
jgi|60181|Punctularia strigosozonata
jgi|123115|Auricularia delicata
gi|323500648|Penicillium decumbens
jgi|121827|Ceriporiopsis subvermispora
gi|194295604|Hypocrea lixii
jgi|164742|Neurospora tetrasperma
jgi|144729|Stereum hirsutum
gi|156059212|Sclerotinia sclerotiorum
jgi|2779|Botrytis cinerea
jgi|63026|Aspergillus sydowii
gi|111256891|Penicillium funiculosum
gi|19455601|Aspergillus oryzae
jgi|196471|Agaricus bisporus
jgi|20383|Chaetomium globosum
jgi|2502|Rhytidhysteron rufulum
gi|170945568|Podospora anserina
gi|83680|Baudoinia compniacensis
gi|148331684|Aspergillus versicolor
gi|163962979|Cryptococcus flavus
jgi|176781|Dothistroma septosporum
jgi|3660497|Claviceps purpurea
jgi|11620|Hysterium pulicare
gi|193230816|Botryotinia fuckeliana
0.1

C)


0.1
gi|188628|Cochliobolus sativus
gi|9112|Setosphaeria turcica
gi|9930|Hysterium pulicare
gi|54456|Trichoderma atroviride
gi|33490476|Talaromyces pinophilus
gi|Trial2|90109|Trichoderma atroviride
gi|4004|Hysterium pulicare
gi|23477178|Chaetomium thermophilum
gi|10479|Coprinopsis cinerea
gi|332629628|Aspergillus niger
gi|38791373|Aspergillus aculeatus
gi|5214|Aspergillus nidulans
gi|37531254|Aspergillus oryzae
gi|759|Botryosphaeria dothidea
gi|66418|Auricularia delicata
gi|163962979|Cryptococcus flavus
gi|148331684|Aspergillus versicolor
gi|83680|Baudoinia compniacensis
gi|176781|Dothistroma septosporum
gi|104286|Setosphaeria turcica
gi|310791716|Glomerella graminicola
gi|11620|Hysterium pulicare
gi|3660497|Claviceps purpurea
gi|137959|Dothistroma septosporum
gi|11125689|Penicillium funiculosum
gi|111013|Hysterium pulicare
gi|15625540|Lentinula edodes
gi|60181|Punctularia strigosozonata
gi|123115|Auricularia delicata
gi|121827|Cerrena unicolor
gi|323500648|Penicillium decumbens
gi|133788|Phanerochaete chrysosporium
gi|76696|Piriformospora indica
gi|104887|Neurospora tetrasperma
gi|76670|Piriformospora indica
gi|109676293|Acrophialophora nainiana
gi|120797|Magnaporthe grisea
gi|41755|Cochliobolus sativus
gi|505261|Humicola insolens
gi|5725372|Setosphaeria turcica
gi|398396678|Zymoseptoria tritici
gi|6123|Stagonospora nodorum
gi|45736|Schizophyllum commune
gi|10725|Coprinopsis cinerea
gi|7658|Coprinopsis cinerea
gi|238599616|Moniliophthora perniciosa
gi|126215798|Leucoagaricus gongylophorus
gi|194295604|Hypocrea lixii
gi|144729|Stereum hirsutum
gi|164742|Neurospora tetrasperma
gi|2779|Botrytis cinerea
gi|156059212|Sclerotinia sclerotiorum
gi|79647|Agaricus bisporus
gi|2502|Rhytidhysteron rufulum
gi|20383|Chaetomium globosum
gi|170945568|Podospora anserina
gi|29461208|Morchella spongiola
gi|19323081|Botryotinia fuckeliana
gi|19456|Aspergillus oryzae
gi|63026|Aspergillus sydowii
gi|10286|Verticillium dahliae
gi|310791835|Glomerella graminicola

D)


0.1
gi|15625540|Lentinula edodes
jgi|83680|Baudoinia compniacensis
gi|148331684|Aspergillus versicolor
gi|163962979|Cryptococcus flavus
jgi|104286|Setosphaeria turcica
jgi|759|Botryosphaeria dothidea
jgi|66418|Auricularia delicata
jgi|76670|Piriformospora indica
jgi|104887|Neurospora tetrasperma
gi|29461208|Morchella spongiola
gi|310791835|Glomerella graminicola
jgi|10286|Verticillium dahliae
gi|238599616|Moniliophthora perniciosa
jgi|10725|Coprinopsis cinerea
jgi|7658|Coprinopsis cinerea
jgi|45736|Schizophyllum commune
gi|126215798|Leucoagaricus gongylophorus
jgi|6123|Stagonospora nodorum
gi|398396678|Zymoseptoria tritici
jgi|76696|Piriformospora indica
gi|505261|Humicola insolens
jgi|120797|Magnaporthe grisea
gi|325372|Setosphaeria turcica
gi|156059212|Sclerotinia sclerotiorum
jgi|2779|Botrytis cinerea
jgi|164742|Neurospora tetrasperma
gi|194295604|Hypocrea lixii
jgi|144729|Stereum hirsutum
gi|193230816|Botryotinia fuckeliana
gi|170945568|Podospora anserina
jgi|20383|Chaetomium globosum
gi|1945601|Aspergillus oryzae
jgi|63026|Aspergillus sydowii
jgi|2502|Rhytidhysteron rufulum
gi|3660497|Claviceps purpurea
jgi|11620|Hysterium pulicare
jgi|123115|Auricularia delicata
jgi|60181|Punctularia strigosozonata
jgi|133788|Phanerochaete chrysosporium
gi|323500648|Penicillium decumbens
jgi|121827|Ceriporiopsis subvermispora
gi|310791716|Glomerella graminicola
jgi|111013|Hysterium pulicare
jgi|137959|Dothistroma septosporum
gi|11125689|Penicillium funiculosum
jgi|79647|Agaricus bisporus
gi|375312547|Aspergillus aculeatus
gi|332629628|Aspergillus niger
gi|387913736|Aspergillus nidulans
jgi|4004|Hysterium pulicare
jgi|5214|Aspergillus oryzae
gi|109676293|Acrophialophora nainiana
gi|Triat2|90109|Trichoderma atroviride
jgi|54856|Trichoderma atroviride
gi|334904676|Talaromyces pinophilus
jgi|9930|Hysterium pulicare
gi|91125|Setosphaeria turcica
jgi|188628|Cochliobolus sativus
jgi|104479|Coprinopsis cinerea
gi|23477178|Chaetomium thermophilum

Détermination de sondes oligonucléotidiques pour l'exploration à haut-débit de la diversité taxonomique et fonctionnelle d'environnements complexes

Résumé :

Les microorganismes, par leurs fascinantes capacités d'adaptation liées à l'extraordinaire diversité de leurs capacités métaboliques, jouent un rôle fondamental dans tous les processus biologiques. Jusqu'à récemment, la mise en culture était l'étape préliminaire obligatoire pour réaliser l'inventaire taxonomique et fonctionnel des microorganismes au sein des environnements. Cependant ces techniques ne permettent d'isoler qu'une très faible fraction des populations microbiennes et tendent donc à être remplacées par des outils moléculaires haut-débit. Dans ce contexte, l'évolution des techniques de séquençage a laissé entrevoir de nouvelles perspectives en écologie microbienne mais l'utilisation directe de ces techniques sur des environnements complexes, constitués de plusieurs milliers d'espèces différentes, reste néanmoins encore délicate. De nouvelles stratégies de réduction ciblée de la complexité comme la capture de gènes ou les biopuces ADN représentent alors une bonne alternative notamment pour explorer les populations microbiennes même les moins abondantes.

Ces stratégies à haut-débit reposent sur la détermination de sondes combinant à la fois une forte sensibilité, une très bonne spécificité et un caractère exploratoire. Pour concevoir de telles sondes plusieurs logiciels ont été développés : PhylGrid 2.0, KASpOD et ProKSpOD. Ces outils généralistes et polyvalents sont applicables à la sélection de sondes pour tout type de gènes à partir des masses de données produites à l'heure actuelle. L'utilisation d'architectures de calculs hautement parallèles et d'algorithmes innovants basés sur les *k*-mers ont permis de contourner les limites actuelles. La qualité des sondes ainsi déterminées a pu permettre leur utilisation pour la mise au point de nouvelles approches innovantes en écologie microbienne comme le développement de deux biopuces phylogénétiques, d'une méthode de capture de gènes en solution ainsi que d'un algorithme de classification des données métagénomiques. Ces stratégies peuvent alors être employées pour diverses applications allant de la recherche fondamentale pour une meilleure compréhension des écosystèmes microbiens, au suivi de processus de bioremédiation en passant par l'identification de tous types de pathogènes (eucaryotes, procaryotes et virus).

Mots clés : *bioinformatique, métagénomique, détermination de sondes, capture de gènes, biopuces, classification*

Selection of oligonucleotide probes for high-throughput study of complex environments

Abstract:

Microorganisms play a crucial role in all biological processes related to their huge metabolic potentialities. Until recently, the cultivation was a necessary step to appraise the taxonomic and functional diversity of microorganisms within environments. These techniques however allow surveying only a small fraction of microbial populations and tend to be consequently replaced by high-throughput molecular tools. While the evolution of sequencing technologies opened the door to unprecedented opportunities in microbial ecology, massive sequencing of complex environments, with thousands of species, still remains inconceivable. To overcome this limitation, strategies were developed to reduce the sample complexity such as gene capture or DNA microarrays.

These high-throughput strategies rely on the selection of sensitive, specific and explorative probes. To design such probes several programs have been developed: PhylGrid 2.0, KASpOD and ProKSpOD. These multipurpose tools were implemented to design probes from the exponentially growing sequence datasets in microbial ecology. Using highly parallel computing architectures and innovative *k*-mers based strategies allowed overcoming major limitations in this field. The high quality probe sets were used to develop innovative strategies in microbial ecology including two phylogenetic microarrays, a gene capture approach and a taxonomic binning algorithm for metagenomic data. These approaches can be carried out for various applications including better understanding of microbial ecosystems, bioremediation monitoring or identification of pathogens (eukaryotes, prokaryotes and viruses).

Keywords: *bioinformatics, metagenomics, probe design, gene capture, DNA microarrays, binning*