Mémoire

Auteur : Debras, Maud
Promoteur(s) : Baurain, Denis; Cornet, Luc
Faculté : Faculté des Sciences
Diplôme : Master en bioinformatique et modélisation, à finalité approfondie
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/15277

LIÈGE université
Sciences

Analysis of secondary metabolite biosynthetic gene clusters in lichen metagenomes

Maud DEBRAS

Master en Bioinformatiques et Modélisation

Finalité approfondie

Année académique 2021-2022

Promoteur : Denis BAURAIN ; Co-promoteur : Luc CORNET

Contents

List of Abbreviations

ANI: Average Nucleotide Identity
BGC: biosynthetic gene cluster
BIOM: Biological Observation Matrix
bp: base pair
BLAST: Basic Local Alignment Search Tool
CDS: coding DNA sequence
COG: Clusters of Orthologous Groups
CONCOCT: clustering contigs on coverage and composition
DNA: deoxyribonucleic acid
EMBOSS: The European Molecular Biology Open Software Suite
PCA: principal component analysis
LBA: label propagation algorithm
HMM: Hidden Markov Models
JSON : JavaScript Object Notation
KEGG: Kyoto Encyclopedia of Genes and Genomes
KO: KEGG Orthology
LCA: Lowest common ancestor
MAG: Metagenome assembled genome
MIBiG: Minimum Information about a Biosynthetic Gene cluster database
MS/MS: tandem mass spectrometry
NCBI: National Center for Biotechnology Information
NGS: next-generation sequencing
NRPS: non-ribosomal peptides
no: nucleotide
PKS: polyketide synthase
RNA: ribonucleic acid
SCG: single-copy core gene
ACOG: Secondary metabolism protein family analysis
SSU: small subunit
TNF: tetranucleotide frequency
UniProt: Universal Protein Resource
WGS: whole-genome sequencing

INTRODUCTION

Natural products are highly important molecules that usually originate from bacteria, fungi, and plants. Also known as secondary metabolites, natural products have been extensively exploited but continue to be an unparalleled resource of bioactive compounds, many of which have found broad applications in medicine, the food industry, cosmetics, agriculture, or ecological bioindication (Zhang *et al.*, 2021). The immense chemical diversity of these compounds stems from the ability of organisms to produce molecules that enable them to cope with various challenges in their environment. These specialized metabolites are involved in survival and interaction mechanisms such as chemical control, nutrient acquisition, or protection against stress (Chavali and Rhee, 2018).

Particularly promising in this context, interest in lichen-forming organisms has increased, with greater awareness of the prospects in the search for natural products. Lichens result from the symbiotic association between a fungus (Ascomycota in 98% of cases) and a photosynthetic partner (mainly green algae, sometimes replaced or accompanied by cyanobacteria) (Grube and Berg, 2009). Due to their dynamic symbiotic profiles, experimental refractoriness, extremophilic habitats, and unparalleled ecological range, lichen symbioses are eminently suitable research objects for bioactivity exploration (Zhang *et al*., 2007). The unique chemo diversity resulting from this particular way of life consists of more than 1000 specialized specialized metabolites (depsides, depsidones, dephenylethers,...). These unique structures found nowhere else in nature, make lichens excellent contenders in the search for new molecules of therapeutic interest (Michal Goga, 2018). However, the bioactive potential of lichens has not yet been fully exploited, due to the limited availability of material from traditional isolation and the difficulty of replicating their chemical arsenal. In light of emerging resistance against prevalent drugs, the growing interest in these molecules is therefore accompanied by the need to overcome experimental challenges using new strategies (Boustie and Grube, 2015; Hengqian *et al*., 2020).

Recent improvements in genomic DNA sequencing techniques have led to new opportunities, allowing the assembly of multi-species metagenomes as well as the study of their enzymatic pathways involved in metabolite production through genome mining. Metagenomics is the study of ecological communities of microorganisms, which aims to assess their population dynamics and functional potential (Zhang *et al.*, 2021). This approach has not only been used to inventory their diversity but has also provided new insights into the multi-species interactions that drive these communities, including complex symbiotic ecology. Metagenome sequencing is a valuable technique for improving knowledge of the genomes of lichen-forming symbionts, offering the ability to obtain complete microbial profiles without prior knowledge of composition, and irrespective of the possibility of culturing them (Ghurye *et al*., 2016; Calchera, 2019). In this case, the individual genomes of the symbiotic partners have to be reconstructed from complex DNA mixtures. However, converting these large volumes of metagenomic data into biologically meaningful information is an extremely difficult, computationally, and methodologically demanding task that remains a major challenge (Zhang *et al*., 2021). The exponential growth of sequencing data has led to the need for ever more sophisticated computational technology to process these valuable resources. Consequently, this need has propelled the increasing evolution of bioinformatics tools. This plethora of tools have emerged to facilitate genome annotations and to handle the ever-increasing size of input sequences and associated databases.

This study highlights the exploration of the biosynthetic potential of six cyanolichens from different Peltigera species for the development of an efficient bioinformatics pipeline in the context of metagenomics. This pipeline presents a solution for the assembly of genomes of interest for genomic exploration from metagenomic data without reference. As the metagenomic data currently in use was retrieved from a previous population genomics study (Magain *et al.,* 2017), the community members in this collection had not yet been screened for biosynthetic relevance. Therefore, the purpose of this genome exploration was to preliminarily assess the value of these untapped cyanobionts in the context of natural product research.

This manuscript is organized into three chapters: an introduction to the symbiotic organisms lichens and the biological context and DNA sequencing technologies needed to process them, a detailed presentation of the tools and algorithms implemented within the methods developed for this work, and finally, the third chapter discloses the presentation of the results and the evaluation of the methods followed to finally discuss the limitations and possible improvements for future research.

PART I: Bibliographic Review

I. Presentation of Lichens

1. Position in the living world

For many centuries the taxonomy of lichens was incorrect, and the dual nature of lichens was only first described in 1867 by Schwendener and De Bary (Honegger, 2000). The symbiosis theory then made it possible to highlight the distinctive identity of these organisms, with great diversity in terms of physiology and taxonomy. Today, it is therefore accepted that lichens represent a single division in the living kingdom of Fungi. Due to the absence of chlorophyll in these organisms, their heterotrophy to organic matter leads them to feed in different ways, including symbiosis which refers to the biological association, sustainable and mutually profitable between different organisms (Gonnet *et al*., 2017). Lichenization, therefore, represents a nutritional strategy, among others, to overcome the heterotrophy of fungi (Farrar J.F., 1976). However, the last decades of research have led to an expansion of the definition of symbiosis and the understanding of lichens that are now seen as self-sustaining micro-ecosystems (Grimm *et al*., 2021; Spribille *et al*., 2022). Modern phylogenetic data suggest that the phenomenon of lichenizationmechanization appeared independently and repeatedly during evolution. This living model is now widespread in one-fifth of the currently known global fungal biodiversity (Lücking, 2019; Zakeri *et al.,* 2022).

2. Definition

A lichen is a reciprocal symbiotic association between a fungal partner and a population of unicellular or filamentous algae or cyanobacteria (Figure 1). The fungal element of the lichen is called mycobiont (from the Greek makes, "mushroom" and *bios*, "life"), and the photosynthetic element is the photobiont (*photo-*, "light" and *bios*, "life"). The name of a lichen species corresponds to the fungal partner, as it alone ensures sexual reproduction. In 98% of cases, the fungus is ascomycetes, the remaining are basidiomycetes. For each species of lichen, there is a corresponding and distinct species of fungi (Lutzoni and Miadlikowska, 2009). Of the ~19 000 species of lichens (Spribille, 2022), only a hundred photobiont species have been reported to be associated with a fungal partner, most commonly the green algae *Trebouxia* and *Trentepohlia*. Moreover, roughly 10% of lichens associate with cyanobacteria as their primary photobiont, termed cyanolichens. While still unclear, the diversity of cyanobionts of bipartite lichens has so far been restricted to the heterocytous phylogroups from the order Nostocales, for which the most common genus in lichen symbioses is *Nostoc* (Rikkinen, 2017).

3. Symbiosis

Cyanobacteria can either be the unique photobiont of the fungus (bipartie symbiosis) or be present as a complement to the eukaryote green alga (tripartite symbiosis) (Rikkinen, 2015). This symbiosis can therefore potentially involve three kingdoms: Fungi, Plantae, and Bacteria. The symbiotic organism is structured by a rudimentary vegetative organism, the thallus, composed of an intimate mixture of fungal hyphae and photobiont (Figure 1). In this perfectly autotrophic structure, the alga or cyanobacteria synthesizes energy-rich organic matter from carbon dioxide (CO_2) in the atmosphere and solar radiation (photosynthesis). While the mycobiont of cyanolichens generally benefits not only from photosynthesis but also from a large uptake of atmospheric nitrogen by cyanobiont; in contrast, the fungus absorbs water and mineral salts from the environment essential to lichen symbiosis. The fungal partner is also responsible for the anchoring of the structure and protects the lichenic association from aggressive ultraviolet radiation and its possible deleterious effects. In this manner, this symbiotic association results in the formation of a distinct lichenic body, which resembles neither of the two partners living in a free state (Rai, 2002; Nash, 2008).

Figure 1. Chlorolichens (*Parmelia sulcata*, A) and cyanolichens (*Lobaria virens*, B) are the two categories of lichens. C- Structure of the lichen thallus. Photobionts are only located under the upper cortex (Scanning Electronic Scan Microscopy of the thallus of *Parmelia sulcata*). [Photos reproduced with the permission of Yannick Agnan].

The traditional characterization of lichens as a dual symbiotic partnership has recently been reconsidered in line with microbiome studies. Indeed, complex interactions between a high diversity of organisms have been established showing the formation of ecosystems harboring less tightly integrated partners with auxiliary functions (Grimm *et al.*, 2021). These interactions with the associated microbiome can notably occur at the metabolic level through exchanges of metabolites (Hawksworth and Grube, 2020). To date, but not well established, this diversity mainly located inside the thallus is called 'endolichenic' (Aschenbrenner *et al.*, 2016). However, research has shown that bacterial communities can be specific to their hosts in different genera of lichenized fungi (Sierra *et al.*, 2020) while the composition of the bacterial communities of related lichens is generally similar (Ochman *et al.*, 2010). The different phyla constituting these communities are most frequently: Actinobacteria, Bacteroidetes, Proteobacteria, and Verrucomicrobia (Sierra *et al.*, 2020).

4. Ecology

Lichens are common across the globe in most terrestrial ecosystems, from the polar regions to the tropics, and account for about 8% of the land cover (Larson, 1987; Asplund and Wardle, 2016). They are found in many habitats ranging from marine or freshwater environments to extreme environments such as deserts and glacial areas and are therefore able to survive in hostile, cold, hot, or even toxic environments. In these extreme niches, photobionts and mycobionts are not present outside these symbiotic associations. Symbiosis is therefore the key to understanding the evolutionary success of these organisms within such ecological habitats (Nash *et al*., 2008).

Lichens are epiphytic organisms growing on the surface of other substrates, in place of having their roots rooted in the soil. Some lichen species prefer to colonize trees, while others choose rocky substrates or soil. Lichens spread in mossy mats and often develop small leaf-like structures that give them the appearance of rudimentary plants. The thallus may carrseveralof different organs: fibrils, spinules, haptens, rhizines, papillae, tubers, cephalods, cyphelles, and pseudocyphelles (Lexicon available in Annex 1). These then aid the identification of the species, using the determination keys. The climatic conditions (sun exposure, humidity, heat) are also important in the establishment of certain species according to their affinities (Van Haluwyn *et al,* 2013).

5. Singular phytochemistry of lichens

This unique symbiotic partnership results in the production of many bioactive natural products, 90% of which are unique to lichens (Stocker-Wörgötter, 2008). These secondary metabolites also referred to as specialized metabolites, emanate from three biosynthetic pathways (Figure 2):

- **Shikimic acid** pathway
- **Mevalonic acid** pathway
- **Acetylpolymalonyl** pathway (polyketide pathway)

The ‘backbone’ or ‘signature’ enzyme synthase and/or synthetase involved in these biosynthetic pathways defines the chemical class of the generated secondary metabolites (e.g. polyketides) which results in approximately twenty structure families (aromatic and aliphatic, the majority of which derive from the acetylpolymalonyl pathway). These products include depsides, depsidones, depsones, dibenzofuranes, anthraquinones, xanthones, chromones, terpenes, diphenylethers, naphthoquinones, steroids, carotenoids, and pulvinic acid derivates, being the principally recognized compounds. These molecules usually have low molecular weights and are essentially produced by the fungal partner, which releases them at the level of the cortex or in the internal structure (Roullier, 2010). The development of secondary metabolites via these three routes of synthesis varies from one species to another but also within the same species. Indeed, a biosynthetic pathway may be more in demand than another in certain environmental conditions. The production of substances may depend on abiotic factors of habitat such as geographic and altitudinal conditions, hydration, chemical signals, UV radiation, climatic fluctuations, and biological factors such as contact with other microorganisms,

predation, or the presence of competing organisms (Calchera, 2019). Although the chemical arsenal is therefore unequally functional from one species to another, major and ubiquitous metabolites can be observed such as atronorin, usnic acid, and stictic acid (LePogam, 2016).

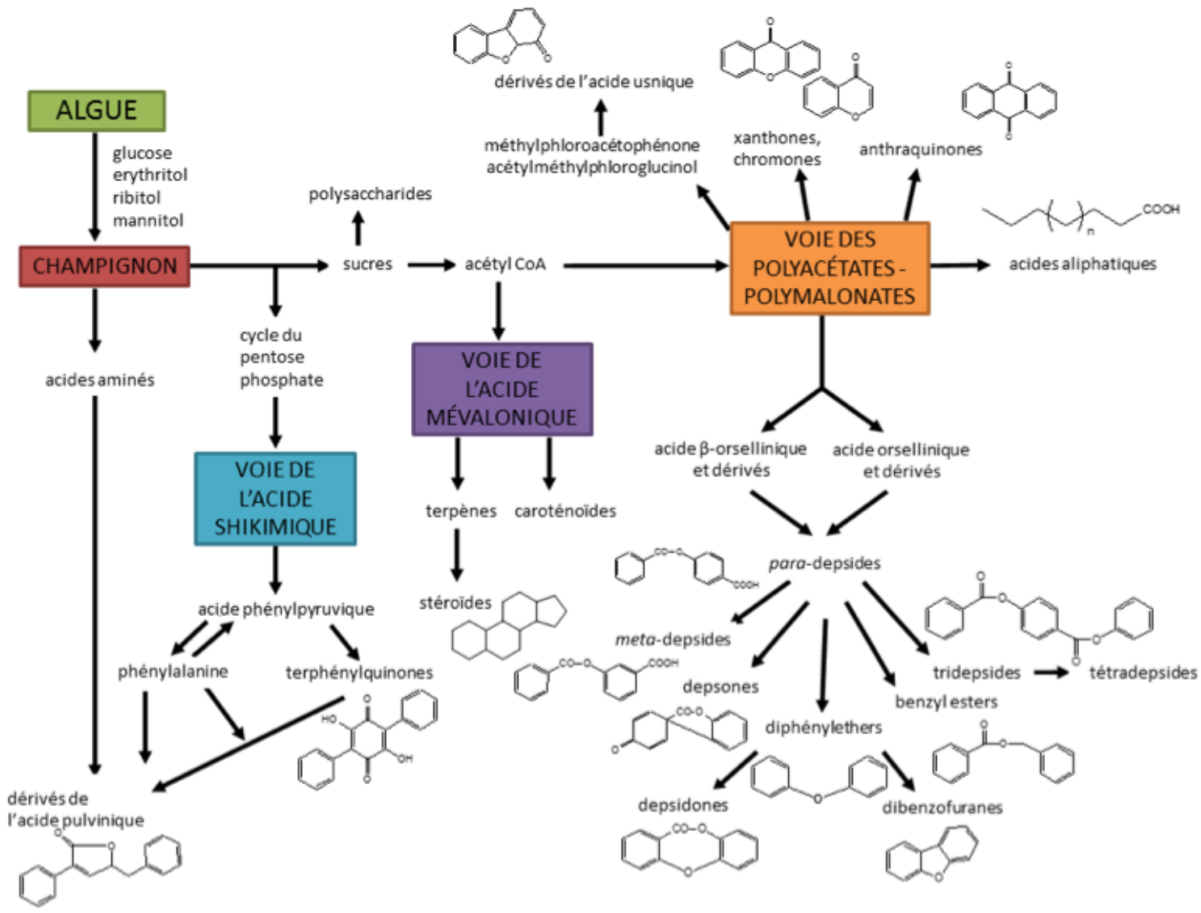


Figure 2. Biogenesis pathways of lichen secondary metabolites (illustrated by some representative structures). The synthesis of secondary metabolites primarily involves the polymerization of primary metabolites by dedicated enzymes (often referred to as backbone or core enzymes). The metabolites generated by the backbone enzymes are further modified by tailoring additional enzymes, that can vastly alter the bioactivity of the metabolites. The backbone enzyme defines the pathway of biosynthesis and accordingly the chemical class of the generated secondary metabolites. The most abundant classes of lichen secondary metabolites are phenolic compounds (e.g. depsides, depsidones, dibenzofurans) built by the enzyme polyketide synthase (PKS) from the acetylpolymalonyl pathway (Keller, 2019). [Figure copy from LePogam, 2015; adapted from Stocker-Wörgötter, 2008].

The number of known secondary metabolites is still increasing, as are the fields of interest in these products. For example, these compounds are used in industry (perfumery, pest control, etc.) or bioindication (air quality and general pollution) (Gonnet *et al.*, 2017). The bioactivities of these molecules are also increasingly studied in therapeutics. Various medicinal properties have been identified, such as antioxidant, antibiotic, antiviral, antifungal, cytotoxic, antipyretic, analgesic, and photoprotective (Boustie *et al.*, 2005). An overview of the described natural products derived from lichens and their biological potential is provided in the review of Goga *et al.* (2018).

6. The lichen genus Peltigera

For the present metagenomic study, data from six species belonging to the genus *Peltigera* (Peltigerineae, Lecanoromycetes) have been made available. *Peltigera* is one of the most widespread lichen genera, especially abundant in boreal and tropical mountain forests. This genus responds to a specific photobiont lineage since all species of lichenized fungi within this genus are associated with a cyanobacterium of the genus *Nostoc* (Vitikainen, 1994). The majority of *Peltigera* species are associated solely with *Nostoc* (bipartite symbiosis), while about 10% of them are also associated with a green alga of the genus *Coccomyxa* in a tripartite association (Miadlikowska *et al*., 2004) (Figure 3). Moreover, in addition to photosynthate delivery within thalli, cyanobacterial symbionts play a crucial role in the arctic and boreal ecosystems of *Peltigera* by their nitrogen fixation capability, particular in these poor environments (Henriksson *et al*., 1971). The relative importance of these activities may vary between different types of cyanolichens (Rikkinen, 2013).

Position in classification:
Peltigeracea, Peltigerales,
Lecanoromycetes, Pezizomycotina, Ascomycota, Fungi

Thalli: large foliose, rarely small to almost squamulose, approximately circular in outline, (1-)2-30 cm in diam.

Lobes: +flattened and elongate (5-15 mm wide and up to 5 cm long), often dichotomously branched, imbricate or separate

Upper surface: gray, bluish gray, grayish brown to brown when dry, bluish gray, blackish green or bright green when wet, smooth, dull or shiny, +scabrid, tomentose or pruinose; with or without isidia or soredia (Soredia: black-brown to orange, most often elongated finger-shaped, and located at the tips of the lobes)

Lower surface: brownish white, densely arachnoid-tomentose or with anastomosing pale or dark brown to black veins, rhizinate; rhizines: white, brown or black, simple, bushy or fasciculate

Substrate: most often on soil and among mosses over rocks, rarely on tree trunks

Geography: world-wide in moist habitats, especially cool temperate ones in North and South America, Europe, Asia, Australasia and Africa

Figure 3. General description of the lichen Genus Peltigera [Asta, Gaveriaux and Van Haluwyn, 2013]. Photography: Bipartite and tripartite cyanolichens. A. Bipartite cyanolichen (*Peltigera scabrosa*, Peltigerales) with cyanobacterial symbionts (Nostoc) in a layer below the upper cortex of the stratified thallus. B. Tripartite cyanolichen (*Peltigera aphthous*, Peltigerales) with cyanobacterial symbionts (Nostoc) in cephalodia on the upper surface of the green algal thallus. C. Tripartite cyanolichen (*Solorina crocea*, Peltigerales) with cyanobacterial symbionts (Nostoc) and green algal symbionts co-occurring in the photobiont layer. Scale bars 1 cm [Figure copy from Rikkinen, 2015].

In comparison with other macrolichens, the taxonomy of the genus *Peltigera*, including their chemotaxonomy, remains poorly understood. In *Peltigera-Nostoc* symbiosis, the specificity of mycobionts in their association with cyanobionts is highly variable. Multiple *Nostoc* species are thought to be involved in the partnership of *Peltigera* and widely intermixed between different hosts (Rikkinen, 2013). There is still some general confusion at the species level taxonomy of cyanobacteria

and symbiotic specificity characterization within the *Peltigera* members. This relates to early species descriptions based on insufficient observations and the unstable nature of cyanolichens that are especially recalcitrant to isolation in culture (Magain *et al.*, 2016). Comprehensive evolutionary phylogenetic studies of the genus are scarce but emerging. The most recent studies have led to the discovery of multiple new species in the revised genus (Miadlikowska and Lutzoni, 2000; O'Brien *et al.*, 2005; Magain *et al.*, 2017a, b; Pardo-De la Hoz *et al.*, 2020) which endorses the expectation of substantial, multiple, undescribed species in the genus *Peltigera*. The genus *Peltigera* also includes species showing different chemical and biological patterns such as antibacterial and antifungal properties (Suleimen *et al.*, 2021). Although a restricted number of lichen compounds have been detected in *Peltigera* members, a large number of mycobiont and cyanobiont genes correlated with secondary metabolites biosynthesis have been identified by metagenomic studies (Grube *et al.*, 2014; Bertrand and Sorensen, 2018; Miadlikowska *et al.*, 2018; Calchera, 2019). The majority of chemical classes reported in the *Peltigera* mycobiont have been depsides and terpenoids produced by the major acetylpolymalonyl and mevalonic acid pathways (Miadlikowska and Lutzoni, 2000). Moreover, among other interesting products, an unusual polyketide biosynthetic pathway has been identified in the *Nostoc* symbiont which synthetizes a unique symbiosis-associated natural product, nosperine, also a promising candidate for bioapplications (Kampa *et al.*, 2013).

This thesis focuses on six cyanolichens from *Peltigera* investigated during recent molecular research that led to the taxonomic characterization of novel species in this genus. This research conducted by Magain *et al.* (2017) showed that lichen-forming fungi previously identified as *Peltigera neopolydactyla* consisted of an assemblage of ca. 10 cryptic species (Magain *et al.*, 2017a, 2017b). The phylogenic tree of the lichen-forming genus *Peltigera* adapted by Magain *et al.* for this mentioned study is shown in Annex 2. Five of these species are closely related and form a monophyletic group with *Peltigera pacifica Vitik*., an endemic species from Northwestern North America. However, none of them corresponds to *Peltigera neopolydactyla sensu stricto,* and they are thus all undescribed. A manuscript formally describing twenty-three new taxa, including these five new species, was recently submitted for publication (Magain *et al.,* in review; submitted to Persoonia in November 2021). In addition to *P. pacifica*, this monophyletic group of six species consists of *P. appalachensis* ad. int. (corresponding to *P. neopolydactyla* 1 in Magain *et al.* 2017a and 2017b), *Peltigera borinquensis* ad. int. (corresponding to *Peltigera neopolydactyla* 1b), *P. vitikainenii* ad. int. (corresponding to *P. neopolydactyla* 2a), *P. mikado* ad. int. (corresponding to *P. neopolydactyla* 2b) and *P. asiatica* ad. int.(corresponding to *P. neopolydactyla* 3). For ease of reading, the present manuscript will refer to their soon-to-be-published name without "*ad. int.*" in the following parts of the work.

II. Genome mining for the discovery of secondary metabolites

1. The context of natural products

Faced with the incredible variety of specialized metabolites offered by nature, many fields have already exploited the wide chemical diversity available through microorganisms (bacteria, viruses, fungi, microalgae) and plants. Of greatest significance, a remarkable number of pharmaceutical drugs (e.g. antibiotics, antifungals, immunosuppressants, *etc*) have been developed, or inspired, by natural products. Recently, the discovery of a new variety of medicines has been highlighted for countering the environmental problems related to synthetic drugs and emerging antibiotic resistance (Van der Hooft *et al.*, 2020). Additionally, microbial natural products are used in many others fields such as the cosmetic, food, and agriculture industries, and are important components for the understanding of ecological interactions. Nevertheless, specialized metabolites found in nature remain a promising source of discovery as more than 80% of estimated biodiversity remains unexplored for biological activities (Sorokina and Steinbeck, 2020).

Traditionally, natural products were isolated from individual microbes and plants and extracted using laborious bioactivity-guided protocols. During the middle decades of the 20th century, this research enabled the reporting of thousands of metabolites and provided the backbones of the majority of therapeutic compounds available today (Wright, 2019). Over time the discovery rate tailed off due to the increasingly frequent rediscovery of known molecules. This headway made was the reason why many pharmaceutical companies turned to the discovery of other leads, such as chemical synthesis pathways. However, over the last thirteen years, natural products have driven the drug discovery of the great majority of relevant molecules. Only 27% of all approved drugs during these decades have been derived from totally synthetic protocols (Newman and Cragg, 2016). This renewed interest in bioactive natural products took place when a successful improvement in sequencing and bioanalytics was achieved. These technologies have together provided a deeper understanding of these resources through screening their genetic makeup, especially for microorganisms, and have demonstrated that their metabolic potential was underestimated (Rokas *et al.,* 2018; Albarano *et al.,* 2020).

2. Genome sequencing technologies for metagenomic

Biomolecular studies have long been restricted to single-species organisms isolated and cultivated purely in the laboratory. Consequently, the highest range of microbial diversity remains inaccessible as only 1% estimated of microbial diversity can be identified by cultivation (Rappé and Giovannoni, 2003; Chi-chu, 2010; Epstein, 2013; Berdy *et al.*, 2017).

The current generalization of new high-throughput sequencing technologies, known as next-generation sequencing (NGS), makes it possible to simultaneously sequence the entire DNA of a sample and thus the sequencing of environmental samples, referred to as the metagenomic study. At the output of a sequencer machine, pieces of DNA sequence called 'reads' have a length that does not exceed a few hundred nucleic bases and can come from any region of the genome. The different high-throughput sequencing technologies currently available stand out on three essential points: the length of the reads obtained, their quantity, and their accuracy. Their main characteristics are summarized in Table 1.

Table 1. Characteristics of the main sequencing techniques. [Table from Goodwin *et al.*, 2016]

	Technology	Length of the reads (bases)	Error rate	average cost per gigabase
First generation	Sanger	400-900	<0.1%	NA
Second generation	Illumina	150 300 (paired reads)	<0.1%	$7 (HiSeq X)
	Roche 454	400	1 %	$9500
	ABI SOLiD	75	<0.1%	$70
Third generation	Pacific Bioscience	10k in average	~5%	$1000
	Oxford Nanopore	10k in average	~5%	$750

Given the complexity and diversity of microbial communities, metagenomics requires significant sequencing efforts and its growth has only been possible due to the democratization of these high-throughput sequencing techniques. Proposed in 1998 by Handelsman, the term metagenomics now refers to environmental genomics, community genomics, ecogenomics, microbial population genomics, etc., where the research angle allows freedom from the constraints of the isolation and culture of the strains, and thereby have access to genetic information on the uncultivable organisms (Bernardo *et al.*, 2013).

Thanks to these revolutionary advances the discovery of natural products is now expanding towards metagenomics through high-throughput direct sequencing of all DNA sampled, regardless of species. This approach, called shotgun, involves non-targeted sequencing of the entire genetic material and makes it possible to highlight the gene content of every member of the microbial community, and thus determine their abilities. Thus, the shotgun method can answer the classic metagenomic questions about the members of microbial communities: "Who are they ? " and "What are they capable of ? ". Nevertheless, as random fragments of the complete genomes are sequenced, a much greater sequencing material is required to cover species in low abundance compared to targeted approaches, thus resulting in a substantially larger amount of data. For this reason, while the method of reconstructing genome sequences directly from metagenomic samples circumvents cultivation, it remains analytically demanding. Indeed, the nature of the data (short fragments of DNA sequences), the complexity of genetic materials (repetitions, genomic duplications, hybridization, horizontal transfers, insertions, transpositions, *etc*. of genetic material), and the similarity of the genomes of nearby species make metagenomes reconstruction particularly challenging and provides draft genomes whose quality depends essentially on the chosen tool. Especially, constraints arising from the

risk of generating chimeric assemblages, which are constructed from sequences originating from different strains (Sangwan *et al.*, 2016; Nurk *et al.*, 2017). However, despite these current limitations, the dedicated tools available today have already led to the resurgence of novel metabolite detection, as detailed by Abarano *et al*. (2020) and Scherlach & Hertweck (2021).

3. Genome mining for biosynthetic gene clusters

A crucial feature of genomics-based natural product discovery is to identify promising biosynthetic pathways among the massive amount of genetic information. While the chemical diversity of bioactive products is immense, their biosynthesis is often based on highly conserved biochemistry. In bacteria and many fungi, all enzymes required for the biosynthesis of a specific compound are encoded into secondary metabolite biosynthetic gene clusters (BGCs). These BGCs are composed at least of core genes, which encode the enzymatic synthase activities, then usually include genes encoding tailoring functions, transporters, and pathway-specific regulatory genes (Keller, 2019). For a certain amount of the major classes of natural products like polyketides (PKS), nonribosomally synthesized peptides (NRPS), ribosomally synthesized and postranslationally modified peptides (RiPPs), alkaloids, and terpenes, such “core” enzymes have been identified and characterized (e.g. Figure 4).

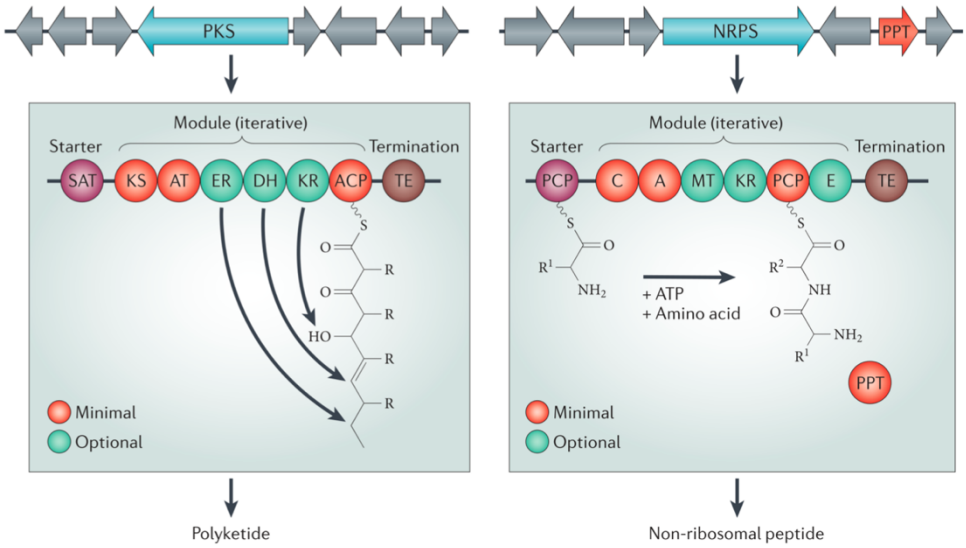


Figure 4. Typical protein domain architecture of the enzyme families of polyketide synthases (PKS) and non-ribosomal peptides synthases (NRPS). Modular PKS and NRPS biosynthetic pathways constitute mechanisms able to assemble complex compounds through the iterative integration of a monomer (amino acids or beta-keto functional groups for NRPS and PKS, respectively) by the adjacent enzymatic modules of the pathway, into a growing chain to elongate their final products. Each functional module of PKS and NRPS comprises specific domains decisive in the selection of which monomers are incorporated. NRPS module contains at least three essential domains: an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, also referred to as the thiolation (T) domain and a condensation (C) domain. Similarly, the PKS module contains at least three catalytic domains: a ketosynthase (KS) domain, an acetyltransferase (AT) domain, and an acyl carrier protein (ACP) domain that are responsible for the selection, activation, and elongation of the growing polyketide chain. Auxiliary modifying domains, like ketoreductase (KR) domains or dehydratase (DH) domains, are found between At and ACP domains (Van der Hooft *et al*., 2020). The family of polyketides synthases is grouped into three different types (I, II, and III PKS) depending on the mechanism of carbon chain elongation, furthermore, the PKS and NRPS pathways can be combined to form hybrid synthetases that form structurally complex molecules, fusing polyketide and amino acid bonds (Calchera, 2019). [Figure copy from Brackhage, 2013].

Therefore, despite the high number of tightly controlled steps comprised in the biosynthetic pathways, the organization of gene coding for these core enzymes in clusters can be leveraged by advanced bioinformatic mining algorithms (Ziemert *et al.*, 2016). Computational approaches dedicated to screening genomes for metabolic pathway prediction are grouped under the term genome mining. Genome mining involves genome assembly and annotation of sequencing data, the recognition of BGCs, prediction of the structure of secondary metabolites, and comparative genomic analysis together with products of the gene clusters identified using experiments (Albarano *et al.*, 2020; Burgos-Toro *et al.*, 2022). Once a BGC is identified, the next challenge is to link it to the corresponding biosynthetic product, to do this most tools rely on similarity comparisons against known functions from public databases. The schematic process of genome mining is illustrated in Figure 5.

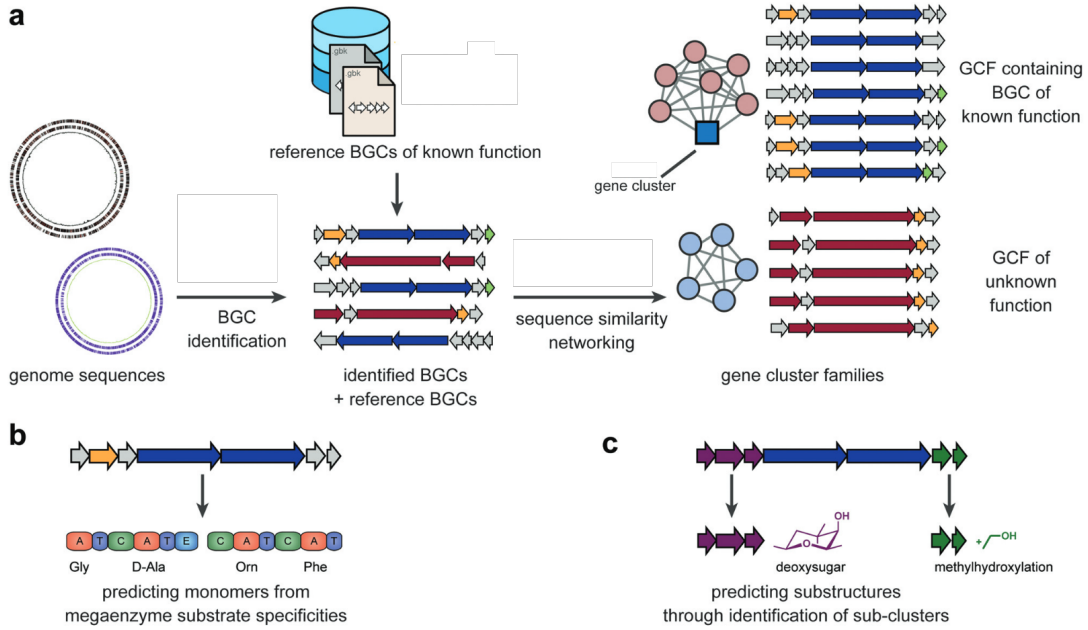


Figure 5. Computational approach processes to mine genomes for metabolic diversity. (a) Biosynthetic Gene Clusters (BGCs) can be automatically identified in genome sequences using bioinformatics tools. Subsequently, they can be dereplicated using databases for BGCs of a known function. Sequence similarity networking can identify groups of similar BGCs across large datasets; the grouping of reference BGCs can aid the annotation of the resulting gene cluster families (GCFs). Two strategies can be employed to predict (partial) chemical structures from these gene clusters: (b) monomers of peptides and polyketides, along with their order, can be predicted using machine learning algorithms for substrate specificity prediction in adenylation (A) domains of nonribosomal peptide synthetases (NRPSs) or acyltransferase (AT) domains of polyketide synthases (PKSs), combined with the analysis of the domain architecture of the whole enzymatic assembly line. (c) Identification of sub-clusters that are known to be responsible for the biosynthesis of specific chemical moieties, or chemical modification, can be used to predict additional structural features of the metabolic product(s) of a BGC. [Figure copy from Van der Hooft *et al.*, 2020].

Due to their typical modular architecture (Figure 4) the enzyme families of polyketide synthases (PKSs) and nonribosomal peptides synthases (NRPSs) are the ones the majority of *de novo* structural predictive tools have focused on (Van der Hooft *et al.*, 2020). Nevertheless, these algorithms are based on current knowledge and thus on the clusters of biosynthetic genes already characterized, and add products with elucidated structure (by using tandem mass spectrometry, *i.e.* MS/MS). This is how, in conjunction with the vast discovery following the effervescence in genome mining, most of the

identified BGCs encode metabolites of unknown structure or function. For these so-called cryptic gene clusters, predicting the structure compound structure is very challenging, albeit hypothetic chemical features may be designated without striving to predict the full structure. Furthermore, given the vast diversity of BGCs that have emerged from metagenomic data, a remarkable number of the predicted BGCs do not match known expressed biosynthetic profiles. Hence, more recently, approaches have been developed to assist in identifying biosynthetic backgrounds. These methods have improved functional inference by including phylogenetic relatedness to characterized genes, transcriptional profiling, or sequence similarity networking that groups functional homologous BGCs into gene cluster families (GCFs) (Wang *et al*., 2018; Kautsar *et al.*, 2020). Today, several of these approaches are often combined to offer more robust putative assignments. An extensive list of currently available tools for genome mining is provided on The Secondary Metabolite Bioinformatics Portal (Weber and Kim, 2016) and many of these have been described in several reviews (Kayani *et al.*, 2021; Baltz, 2021; Van der Hooft's *et al.,* 2020; Chavali and Rhee, 2018; Ziemert *et al.*, 2016).

III. Lichen Genomics

1. Opportunities

Symbiotic relationships shape the functioning and evolution of all organisms but remain incompletely described, in part due to the difficulty in characterizing the genomic diversity that members of the symbiosis collectivity, termed holobiontes (i.e. host and other species living around assembly), are constituted of (Simon *et al.*, 2019). The new genomic area has revolutionized the study of these systems, offering many new opportunities in the context of lichen research, among which particular attention is paid to their abundant secondary chemistry. Indeed, as the poor cultivability of lichen symbionts makes their genomes inaccessible to standard practices, metagenomic approaches constitute especially valuable tools to access their genomic features. The recent results of a widely conducted whole-genomes study on Fungi (Gerasimova *et al*., 2022) revealed a particularly high biosynthetic gene richness harbored by the genomes of lichenized-fungi in the Lecanoromycetes when compared to other fungal classes. The diversity identified through these metagenomic approaches exceeds by far the number of reported lichen-forming fungi metabolites, which highlights the potential production of many unknown metabolites yet to be detected. However, biotechnological applications remain limited since only a few genomes and biosynthetic studies are still available to date (Michal Goga *et al.,* 2018).

2. Knowledge

Lichen symbiosis is an example of multi-species symbiotic association, which is beginning to be more clearly understood thanks to the advent of high-throughput metagenome sequencing technologies. Metagenomic lichen samples reveal ecological and evolutionary aspects of lichen symbiosis that were previously regarded as inaccessible. Over the past decade lichen metagenomics has been employed to study gene expression (e.g. Miao *et al.*, 2012), transcriptome characterization (e.g. Junttila and Rudd, 2012), phylogenetic reconstruction (e.g. Grewe *et al.*, 2017), genomic signatures of adaptation (e.g. Dal Grande *et al.,* 2017), genetic population structure (e.g. Allen *et al.*, 2018) and increasingly, to aid the discovery of biosynthetic pathways in various lichen-forming symbionts. Simultaneously, automatic methods and prediction computations (*in silico* methodologies) have become a principal step of targeted genome mining in lichen research due to the lack of existing knowledge in the literature. Nowadays, approaches applied to lichen genomics integrate the latest DNA technologies and further combine multi-omics analyses to overcome the challenges of working with lichen cultures (Berttrand and Sorensen, 2018; Palazzotto and Weber, 2018). For example, the first identification of a gene cluster in a lichen and its corresponding secondary metabolite, the grayanic acid (lichen-specific depsidone), was accomplished by Armaleo *et al.* (2011) in *Cladonia grayi*, through a combination of transcriptome data with phylogenetic and biosynthetic predictions. In 2012, Kaasalainen *et al.* identified the microcystin-producing gene clusters from cyanolichens by bioactivity screening and amplification of the genes encoding the core enzyme. More recently, metabolic engineering has

enabled the first successful expression of a lichen biosynthetic gene, reported by Kealey *et al.* (2021) to link the PKS cluster from *Pseudevernia furfuracea* to the lichen product, lecanoric acid. To date, the link between a lichenic secondary metabolite and the encoding biosynthetic cluster has been assigned to grayanic acid (Armaleo *et al*., 2011), usnic acid (Goga *et al.*, 2020), lecanoric acid (Kealey *et al*., 2021), physodic acid, olivetoric acid (Singh *et al.*, 2021) and atranorin (Kim *et al.*, 2021).

Furthermore, metagenomic natural product discovery in lichen communities is now increasingly extended to other hyper-diverse partners whose implication in the production of symbiosis-specific natural products continues to amaze (e.g. Kampa *et al.*, 2013; Cernava *et al*., 2017; Leao *et al.*, 2020; Ponsero *et al*., 2021). Of note, the photobionts of cyanolichens are a prolific, but largely untapped source of natural products. The filamentous cyanobacteria *Nostoc* strains often harbor a high amount of BGCs in their genomes, such as PKS and NRP clusters (Shih P. *et al*., 2013). The wide spectrum of bioactivities associated with this biosynthetic potential (e.g., antitumoral, antibiotic, anti-inflammatory, antiviral, and antifungal properties) makes *Nostoc* strains also strong candidates for biotechnological applications (Demay *et al*., 2019; Zahra *et al*., 2020). The biosynthesis of clusters linked to cyanobiont products, such as microcystin and nosperin, is becoming meaningfully understood (Gagunashvili *et al.*, 2018; Dehm *et al.,* 2019).

3. Aims of the study

Due to the slow growth rates of lichens and the difficulties to cultivate them in the laboratory, only a few specimens have been largely exploited for their bioactive potential. Consequently, very few genomes of lichen-associated organisms are presently available while many experimentally recalcitrant species, such as *Peltigera* species, are still poorly studied. To highlight these prolific, but largely untapped sources of new natural products, the aim of this study was the analysis of the genetic richness of the lichen-thalli from six *Peltigera* species without the need for a reference genome.

Taking advantage of advances in high throughput sequencing and metagenomics, the objectives of this work were:

- Design a bioinformatic workflow to assemble draft genomes from shotgun metagenomic data without the need for a reference model
- Refine metagenomic assembly to select metagenome-assembled genomes (MAGs) to appoint candidates to genome mining for biosynthetic potential
- Curate a catalog of tools that can be used and submitted in this workflow taking to account their respective advantages and disadvantages
- Highlight the biosynthetic potential of lichen-associated members from communities related to six untapped *Peltigera* species using *in silico* genome mining technologies

PART II: Materials and Methods

I. Data collection and sequencing

In the context of Next-Generation Sequencing, the most efficient way to sequence a substantial piece of DNA involves a process known as shotgun sequencing, which could be proceeded by the use of Illumina platforms. This technology yields high sequence coverages and detects low abundance members of complex environmental samples.

For this study, metagenomic data of six *Peltigera* species (*P. appalachenis*, *P. asiatica*, *P. borinquensis*, *P. mikado*, *P. pacifica*, *P. vitikainenii*) were provided by the Department of Biology, Ecology, and Evolution (BEE) of the University of Liege. The cyanolichens had been previously collected from multiple harvests across the northern hemisphere for the majority of them, as detailed in Table 2. Genomic DNA was isolated from whole lichen thalli lacking any visible symptoms of fungal infection following two extraction protocols: Cubero et al. (1999) or modified Zolan and Pukkila (1986) using a 2% sodium dodecyl sulfate (SDS) as the extraction buffer. DNA 150 bp paired-end library preparation (DNA-seq, 500 bp insert), and metagenomic sequencing on the platform Illumina NovaSeq 6000 were performed by the Sequencing and Genomic Technologies Shared Resource, Duke Center for Genomic and Computational Biology, Duke University (Durham, North Carolina, USA).

In this approach whole metagenomic DNA is randomly sheared into short fragments of 150 bp, the defined size, then ligated by 5’ and 3’ adapters. These inserts are then both ends sequenced to form thousands of short paired-end reads. Raw reads sequencing data are finally stored in two readable files corresponding to forward and reverse reads respectively. This technique is faster than the hierarchical clone approach but requires high computational and precise algorithms to assemble such an amount of data with high accuracy, particularly when characterizing environmental components.

II. Bioinformatic pipeline

Although numerous benchmarking studies have attempted to find the optimal workflow to process and analyze raw data from sequencing experiments, there is currently no agreed optimal pipeline. In this way, the objectives of this work involved the development of a bioinformatic pipeline that allowed the manipulation of *Peltigera* genomic data.

The procedures and bioinformatics tools followed by this metagenome mining survey were *(i)* the metagenomic assembly of the raw reads achieved by metaSPADES; *(ii)* the binning of assembled sequences into taxonomic compilation using two distinct software: CONCOCT and MetaBAT2; *(iii)* selection of the symbiont-representative metagenome-assembly genomes (MAGs) through quality assessments provided by CheckM, EukCC, BUSCO, and DASTool; *(iv)* functional annotations of genes achieved using MG-RAST; followed by *(v)* the identification of biosynthetic gene clusters involved in the production of secondary metabolites using AntiSMASH.

[The laboratory notebook containing some of the main uses used for these tools and test scripts is available in the public directory: https://github.com/Maud-Debras/Master_Thesis_Lichens-2022]

Table 2. List of the sequenced samples and associated metadata.

Sample	Species	Geographic Origin	Harvest Date	Herbarium	Collector	Concentration Of DNA (ng/µL)
S14	*Peltigera pacifica*	Canada:British Columbia	2012	UBC	Trevor Goward	15
S15	*Peltigera pacifica*	USA:Oregon	2005	OSC	Bruce McCune	12
S16	*Peltigera pacifica*	Canada:British Columbia	1994	H	Orvo Vitikainen	63
S17	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	38
S18	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	60
S19	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	76
S20	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	28
S21	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	14
S22	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	34
S23	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	68
S24	*Peltigera vitikainenii*	Japan:Hokkaido	2019	LG	Magain et al.	128
S25	*Peltigera vitikainenii*	Japan:Honshu	unknown	Tartu	Inga Juriado	30
S26	*Peltigera vitikainenii*	Japan:Honshu	unknown	Tartu	Inga Juriado	111
S27	*Peltigera vitikainenii*	Russia:Khabarovsk	2018	MHA	L Konoreva	58
S28	*Peltigera vitikainenii*	Russia:Sakhalin	2013	SAKH	S. Tchabanenko	66
S29	*Peltigera vitikainenii*	France: Vosges	2019	LG	Magain	58
S30	*Peltigera vitikainenii*	Suisse	1998	G	Mathias Vust	71
S31	*Peltigera vitikainenii*	Canada:British Columbia	unknown	unknown	unknown	25
S32	*Peltigera vitikainenii*	Canada:British Columbia	unknown	unknown	unknown	25
S33	*Peltigera vitikainenii*	USA:Alaska	2013	OSC	K. Spickerman	44
S34	*Peltigera vitikainenii*	Estonia	unknown	Tartu	Inga Juriado	51
S35	*Peltigera vitikainenii*	Norway	2011	LG	Magain	24
S36	*Peltigera vitikainenii*	Norway	2011	LG	Magain	37
S37	*Peltigera vitikainenii*	Norway	2011	LG	Magain	153
S38	*Peltigera vitikainenii*	Norway	2011	LG	Magain	50
S39	*Peltigera appalachensis*	USA:Vermont	2019	LG	Magain Balthazar	95
S40	*Peltigera appalachensis*	USA: North Carolina	2019	LG	Magain Balthazar	63
S41	*Peltigera appalachensis*	USA: North Carolina	2019	LG	Magain Balthazar	142
S42	*Peltigera appalachensis*	USA:Vermont	2019	LG	Magain Balthazar	30
S43	*Peltigera appalachensis*	USA: Alabama	2019	LG	Magain Balthazar	91
S45	*Peltigera appalachensis*	USA: Connecticut	2019	LG	Goffinet	126
S46	*Peltigera appalachensis*	USA: Alabama	2019	LG	Magain Balthazar	81
S47	*Peltigera appalachensis*	USA: South Carolina	2019	LG	Magain Balthazar	80
S48	*Peltigera appalachensis*	USA: Arkansas	2019	LG	Magain Balthazar	91
S49	*Peltigera appalachensis*	USA: North Carolina	2019	LG	Magain Balthazar	22
S50	*Peltigera appalachensis*	USA: North Carolina	2019	LG	Magain Balthazar	13
S51	*Peltigera appalachensis*	Canada: Quebec	2019	DUKE	Miadlikowska et al.	12
S52	*Peltigera appalachensis*	Canada: Quebec	2019	DUKE	Miadlikowska et al.	51
S54	*Peltigera appalachensis*	Canada: Alberta	unknown	unknown	unknown	45
S55	*Peltigera appalachensis*	Canada: British Columbia	unknown	unknown	unknown	16
S56	*Peltigera appalachensis*	Russia: Sakha	2011	H	Teuvo Ahti	54
S57	*Peltigera appalachensis*	Russia: Komi	2019	Tartu	Inga Juriado	95
S58	*Peltigera appalachensis*	Norway	2011	LG	Magain	56
S59	*Peltigera appalachensis*	USA: Alaska	2006	CONN	Bernard Goffinet	74
S60	*Peltigera appalachensis*	USA: Alaska	2011	OSC	Bruce McCune	98
S61	*Peltigera asiatica*	Malaysia: Borneo	2018	Duke	Magain et al.	42
S62	*Peltigera asiatica*	China: Yunnan	2003	Duke	Miadlikowska	16
S63	*Peltigera asiatica*	Taiwan	NA	LG	Serusiaux	9
S64	*Peltigera mikado*	Russia:Khabarovsk	2013	Duke	Miadlikowska	116
S65	*Peltigera mikado*	China: Yunnan	2010	CONN	Goffinet	32
S66	*Peltigera mikado*	Taiwan	unknown	LG	Serusiaux	73
S67	*Peltigera borinquensis*	Peru	2011	Duke	Gaya	41
S68	*Peltigera borinquensis*	Puerto Rico	unknown	Field Museum	Mercado-Diaz	46
S91	*Peltigera appalachensis*	USA:South Carolina	2019	LG	Magain Balthazar	25
S93	*Peltigera appalachensis*	USA:West Virginia	2019	LG	Magain Balthazar	11

All samples mentioned in Table 2 are DNA-Seq (~500bp insert) type. Each sequencing launch was made on full plates (96 samples) with 20µl volume. Sample IDs were automatically suffixed by 6126 code. This table also includes some metadata concerning the sample set such as location and year of harvest. Specimens kept within the herbarium ULiège are by LG.

1. Metagenomes construction

1.1. Quality Control and Trimming

Illumina short-reads were encoded in the FastQ format, which is an ASCII text-based readable format with incorporated base-calling scores. The first step of the bioinformatic pipeline was meant to pre-process data coming from the high throughput sequencing, in preparation for forwarding analysis. To this end, the raw reads were trimmed based on content and stripped adapters with a default setting of *FastP v. 0.19.6* (Shifu *et al.*, 2018), ensuring that only high-quality sequences with a minimum PHRED 15 score threshold were left. This tool aims to provide an all-in-one modular set of analyses to perform quality checks and filtering. Adapters are automatically detected with no need for a supplementary list in the majority of Illumina sequencers, including NovaSeq 6000. For each input file, *FastP* has created a JSON file and an HTML file that reports visual summary statistics under several parameters such as per-base quality of reads, trimmed adapters, duplication rate, and the distribution length of the reads. Appendix X shows the FastP results of the total number of reads, the duplication rate, as well as the percentage of reads discarded during quality trimming.

Finally, *FastQC v. 0.11.9* (Brown *et al.,* 2017) was used to generate quality reports of the raw and final read libraries to assess read quality improvement.

1.2. Metagenomic assembly

Genome assembly refers to the process of bringing together small DNA fragments produced by modern sequencing technologies to obtain fully restored genome sequences, which is either still a challenging process. In this manner, the assembly of mixed samples with many species in different abundances, as is necessary for metagenomics data, remains even more complicated due to (i) the vast data produced, (ii) the quality of sequencing, (iii) the variability of reading coverage between genomes, due to the unequal representation of the community members, (iv) the presence of several strains of a single species, (v) the presence of closely related species that share conserved genomic regions (Nurk *et al.*, 2017, Quince *et al*., 2017, Lapidus *et al.,* 2021).

a. Assembly using MetaSPAdes

To meet these demands, various assembler software has been supplemented with metagenomic-specific extensions. Among the most commonly used, m*etaSPAdes* (Nurk *et al.* 2017, Quince *et al.* 2017) were designated in this study to achieve assembly of the 54 lichen-forming fungi metagenomes. Apart from the *SPAdes* toolkit, this program is a *de novo* assembler based on the de Bruijn graphs approach. *De novo* assemblies assume no prior knowledge of the source DNA sequence length, layout, or composition. This refers to the sequencing when no reference genome is available for alignment, which is common when dealing with complex metagenomics data.

MetaSPAdes was run with the flag *meta* for metagenomic samples, and the Khmer range of 21, 33, and 55 for iterative de Bruijn graph construction was determined automatically based on reading length and sequence data type.

This process first constructs de Bruijn graphs by reading the consecutive *k*-mers within each read. The sliding window is shifted along all sequences, extracting the fixed-size nucleotides. These become the nodes, which are linked together by directed edges. This first step is followed by various graph simplification strategies to transform it into an assembly graph and constructs the paths corresponding with contiguous sequences, so-called contigs, within the genomes sequenced (Vollmers *et al.,* 2016; Nurk *et al.*, 2017). *MetaSPAdes*'s specific strategies were developed to improve the *SPAdes* tool by introducing a novel heuristic approach to dispense with large metagenomics data sets while simultaneously dealing with low coverage and inter-species repeats. An important particularity is how strain variation is handled. The hypothetical *k*-mers of highly similar strain fragments are combined to form quality consensus sequences, which are often longer than contigs from other assemblers.

Finally, *metaSPAdes* generates an assembly graph in GFA format that stores information needed to visualize the assembly. The best-assembled results are then displayed in two FASTA files containing sets of contigs and scaffolds. While contigs are contiguous sequences made from assembled reads, the scaffolds are the result of a subsequent process in which the contigs are ordered and linked with 'N' gaps to provide longer segments.

b. Assembly evaluation

During assembly, each stage must be assessed for quality control purposes to ensure progress. An assembly report was generated using *QUAST v 5.1* (Mikheenko *et al.,* 2018), which evaluates basic statistics. Commonly used quality metrics are total size, several contigs/scaffolds generated, and weighted median contigs/scaffolds size (N50). These metrics such as the number of scaffolds, the total length of metagenomes, the length of the largest scaffolds, the G-C content, the N50/N75 values, and the number of gaps 'N' were collected by *QUAST* and summarized HTML reports.

c. Mapping reads

To guide assembly data, computational alignment of the reads back to sequence was carried out to collect mapping information and generate files expected by downstream bioinformatics analysis. Quality backward and forward reads were mapped against the metaSPAdes assembled scaffolds with *BamM v1.7.3* and *-make* command. This operation converted sequencing data into binary BAM format to store reads as well as their genomic coordinates. Finally, BAM files were sorted based on scaffold length, then indexed using, respectively, *-sort* and *-index* options from *Samtools v 1.3.1* commands.

2. Taxonomic assignment of the reads

For the present study, one sample was chosen per *Peltigera* species to investigate the taxonomic profiling of the metagenomic reads. Six samples were selected based on the maximum scaffolds total length generated by the assembly, assuming that the higher the length of the assembly the more complete the metagenomic data set (Annex 6). The read sets of the samples S59, S61, S68, S64, S14, and S23 were designated for the taxonomic analysis of *P.appalachensis*, *P.asiatica*, *P.borinquensis*, *P.mikado*, *P.pacifica,* and *P.vitikainenii*, respectively.

To proceed with taxonomic classification and abundance profiling, the classifier tool *Kraken* uses *k*-mer signature alignment to associate nucleotide substrings with the lowest common ancestor (LCA) taxa. A genome database is first precomputed according to a default or user-customized references set, then broken into exact signatures (21, 25, 31 nt). For each query read the *k*-mers are organized in a taxonomic tree found by retaining taxa associated with matching signatures passing through phylum to species ranks, depending on the conservation of their component *k*-mers. A read is then assigned based on its composition in k-mers and, more precisely, it is assigned to the most recent common ancestor of the taxa designated by the tree (Lindgreen *et al.,* 2016; Yang *et al.*, 2021).

In the present study, *Kraken v.2* (Wood *et al.*, 2019) was run on the trimmed read pairs with *-username*, *-paired,* and *-report* optional parameters. The specific database required for the program was downloaded from the NCBInr database. The visualization of taxonomic abundances of the results was conducted using *Krona v 2.7.1* (Ondov *et al*., 2011), which offers data to be explored with space-filling displays and multi-layered pie charts. In addition, the Kraken-*biom* package (https://github.com/smdabdoub/kraken-biom) made it possible to extract the standard BIOM formats from the Kraken report outputs to allow further manipulations (e.g. on *R Studio*).

3. Genomes Binning

While assembling is the task of transforming a multitude of short reads into longer proportions of the present genomes, the binning method aims to group sequences of the same taxonomic origin. In the context of the metagenomic project, the majority of current assemblers do not lay out complete microbial genomes with single scaffolds. Additional tools are therefore necessary to find contigs from the same species. Binning tools aim to classify contigs with similar characteristics into organism-specific groups, so-called 'bins'. These resulting bins are commonly known as metagenome-assembled genomes (MAGs) since, ideally, one bin will be generated per unique genomic taxon present in the sample. This process aids the understanding of microbial population structure in addition to the analysis of each species individually inside the sample (Vijini *et al.,* 2020-2021). The main methodologies to perform this critical task are given in Annex 3.

3.1. Binning DNA sequences using two distinct hybrid strategies

To represent the whole genomes of organisms sampled in the various *Peltigera* microbiomes, the segregation of the individual organisms was performed by *CONCOCT v 1.1.0* (Alneberg *et al.*, 2014) and *by MetaBAT v 2.15* (Khang *et al.*, 2019). Both are taxonomic-independent tools that combine sequence composition and differential read coverage to perform binning.

a. Clustering with CONCOCT

CONCOCT combines short *k*-mer frequencies (4-6 nt) and differential abundances to classify the scaffolds using principal component analysis (PCA), and finally group them by fitting a Gaussian mixture model to describe the entire dataset (Sangwan *et al.*, 2016; Roumpeka *et al.,* 2017; Mallawaarachichi *et al.*, 2020; Yang *et al.*, 2021). In this study, the metagenome assemblies of lichens were analyzed with *CONCOCT v 1.1.0* following the recommended protocol with default values: three *k*-mer sizes of 4,5 and 6, a length threshold of 1000 nt, and several groups of 400 for the clustering PCA based on *k*-mer frequencies. Briefly, the metagenome sequences of the assembly scaffolds were cut into non-overlapping segments of 10 kb (*cut_up_fasta.py*) and coverage depth was calculated using the BAM mapping files (*CONCOCT_coverage_table.py*). These coverages were then used during the execution of *CONCOCT* to bin the segments before the resulting clusters were merged (*merge_cutup_clustering.py*) to obtain the final MAG bins (*extra_fasta_bins.py*) (Alneber *et al.*, 2014; Maguire *et al.,* 2020).

b. Clustering with MetaBAT2

Meanwhile, *MetaBAT2* uses tetranucleotide composition information and read coverages to calculate scaffold similarities and build a graph using scaffolds as vertices and their similarities as the weights of the edges. The graph is further partitioned into subgraphs, or bins, by applying a modified label propagation algorithm (LPA) (Soueiden and Nikolski, 2016; Khang *et al.*, 2019). As the same inputs as *CONCOCT*, *MetaBAT2* required assembly scaffolds as well as both BAM mapping files of the read pairs but necessitates no additional settings since *MetaBAT2* was introduced to adapt to the given data to find the best parameters.

3.2. Validation of the binning

Subsequent inspection of the sequence collections resulting from the binning processes is necessary to estimate the reliability of the assemblies obtained, and thus validate the raw material to be used for genome mining analyses. Again, in this study, the assembly statistics, such as the total size of the bins and the number of scaffolds they contain, were calculated by *QUAST v 5.1*. Additionally, clustering accuracy in organism-specific bins was inspected to lead to the validation of binning data.

Generally, although ensuring the consistency of the bins remains difficult, two metrics have been widely adopted to assess the quality of the metagenomic bins. These metrics are estimated based on the presence or absence of lineage-specific genes, from single-copy core gene (SCGs) repertories. The first is completeness, which estimates how completely a metagenomic bin represents a full genome by the percentage of expected single-copy genes that are found in a given bin. The second quality of a metagenomic bin is contamination, which indicates the amount of sequence that does not belong to this population but comes from another genome. This trait is based on the redundancy measured by the number of each SCG found within a bin. Since these genes are hoped in a single copy within a genome, the contamination is estimated from the percentage of single-copy genes found in duplicate. The designation of metagenome bins as MAGs relies on these two parameters, according to classification of high-quality (completeness >90% and contamination <5%) and medium quality (completeness >50%, contamination < 10% and completeness – [5 x contamination] > 10%) (Yang *et al.*, 2021).

As cyanolichens are mixtures of species from different kingdoms (including fungi and cyanoprokaryote symbiont partners), two different tools were designated to estimate the quality of the genomes supposed to originate from Procaryota and Eukoryota, distinctively. These tools were used to assess the completeness and contamination of the bin sets resulting from *CONCOCT* and *MetaBat2*.

a. Completeness and contamination assessed using CheckM

CheckM v 1.1.3. (Parks *et al.,* 2015) uses universal marker genes that are specific to a genome-based lineage to automatically place the bin within a reference tree. Then, it estimates the completeness and the contamination level by checking the sequence for marker genes that a genome of the bin's taxonomy is expected to have, from the bacterial and viral SCG collection provided with the software (Parks *et al.,* 2015). The automatic workflow according to the *lineage_wf* option was followed.

b. Completeness and contamination assessed using EukCC

EukCC (Saary *et al.*, 2020) is an unsupervised estimator specifically developed to assess the quality of novel metagenomic-assembled microbial genomes of eukaryotes. Identical to *CheckM*, this tool is based on the use of a reference set to perform an initial taxonomic classification of the bin, to enable placement in the pre-computed reference tree. Additionally, *EukCC* offers a new reference list enriched with marker genes specific to eukaryote phylogroup. To once again evaluate the completeness and contamination of the bins, *EukCC v 2.0/0.3.* was used with default parameters and the *version 1.1* database.

c. Taxonomic assignment

Finally, the taxonomic assignment sought to dissociate the many components of the symbiosis and characterize the identity of the draft genomes binned. Considering the rich bacterial diversity expected within the lichen environments, the GTDB toolkit specially intended for the taxonomic classification of bacterial and archaeal genomes was chosen (Parks *et al.*, 2018, 2019). The Genome Taxonomy Database (GTDB) is an online reference bank constructed from RefSeq and Genbank genomes. This gold standard taxonomy for procaryote classification uses the criteria of relative evolutionary divergence (RED) and average nucleotide identity (ANI) for establishing rank-normalized taxonomy, which improves the classification. During the present work, the bins were submitted to *GTDB-Tk*

(Chaumeil et al., 2020) which provided taxonomic assignment for draft genomes by placing them into domain-specific, concatenated protein reference trees. Results reported in this study are based on *GTDB-Tk v 1.7.0* and GTDB R06-RS202 (2020).

The complete results of the binning, from both *CONCOCT* and *MetaBat2*, as well as their completeness, contamination, and taxonomic assignments, were reported in a spreadsheet for each of the 55 samples analyzed.

3.3. Selection of the MAGs

To improve the overall quality of the draft genomes the low-quality bins were discarded through manual and automated curation. For this step, the outputs predicted by *CONCOCT* and *MetaBat2* binning algorithms were aggregated by the automated *DASTool* scheme (Darlin *et al.*, 2014) such that the best bins from each binning tool were picked in a resulting consensus set of higher-quality bins. *DASTool v1.1.1* was run on all two collections, using parameters *--search_engine diamond* and *--score_threshold 0.7*. The bins in each collection were compared by *DASTool* and a non-redundant collection of bacterial bins with completeness >70% was created. To complete the refined collection, the range of bins associated with fungi lineage with completeness >70% was manually curated and the best fungal bins were identified based on scores calculation. The scores were calculated according to the F1-score function, which is the harmonic conciliation between precision and completeness:

$$F1 - score = \frac{2 \times \left(\frac{precision \times completeness}{precision + completeness}\right)}{100} \qquad with \;\; precision = 100 - contamination$$

Additionally, for each of the mycobionts and cyanobionts sectioned by the refined collection, quality metrics were again estimated. *BUSCO* (Simão *et al.*, 2015) was used in 'genome mode' with *–the auto_lineage* option, which automatically attempts to find the most appropriate SCMG dataset to use based on phylogenetic placement. An additional option was run during the submission of the fungal bins to request the *AUGUSTUS* algorithm (Stanke *et al.*, 2004).

The bins from the refined collection were one last time for intention to downstream analyses. To avoid a false conclusion, only high-quality bins were retained for the selection of the MAGs. As criteria for inclusion as MAG, the genome had to have greater than 90% completeness, less than 5% contamination, and less than 10% fragmentation. As part of this study, the refined dataset was finally trimmed to contain high-quality MAGs identified as the *Nostoc* cyanobiont to compose subjects for functional assignment and BGC detection.

4. Genome mining

4.1. Functional annotation

Once the primary question of which microorganisms inhabit the environments sampled has been dealt with, the focus usually shifts to what it is that they can do, i.e. the functional potential of each microbial community. As with taxonomic classification, functional annotation is at its most substantive the process of identifying coding regions within sequenced genomes, and aligning these to a translated protein database to carry out predictions based on similarity searches. The open-source server *MGRAST*, short for Metagenomic Rapid Annotations using Subsystems Technology (Meyer *et al.*, 2008), offers an automated pipeline to perform both taxonomic and functional analyses using a dedicated comparative genomic framework. The search is computed against a unique curated databank of genomes (*M5NR*, Wilk *et al.*, 2012), which provides non-redundant integration of sequences from the SEED (Overbeek *et al.*, 2005), GenBank (Benson *et al.*, 2008), RefSeq (O'Leary *et al.*, 2016), KEGG (Kanehisa and Goto, 2000), UniProt (The UniProt Consortium, 2021), IMG (Markowitz *et al.*, 2012) and eggNOGs (Muller *et al.*, 2010) databases and associated annotations (Keegan *et al.*, 2016; Meyer *et al.*, 2019). To contemplate the variety of the biological roles implemented by the sampled cyanobionts, the *MG-RAST version 4.0.3* web server was run on the *Nostoc* MAGs selected with the default settings (annotation source m5NR, maximum E-value cutoff 1e-5; minimum identity cutoff 60%; minimum alignment length cutoff 15). The running time was between 10 to 25 minutes per genome in the datasets. Finally, comparisons of taxonomic and gene function assignments were visualized using a variety of formats, including stacked bar charts, heat maps, and circular charts.

Table 4. List of the databases recruited by the annotation tool MG-RAST

Database	Description	Refrence
M5NR	The M5NR is an integration of many sequence databases into one single, non-redundant, searchable protein database.	(Wilk *et al.*, 2012)
SEED	The SEED database is a collection of subsystems (collections of functionally related protein families) and their derived FIGfams (protein families).	(Overbeek *et al.*, 2005)
GenBank	The Genetic sequence database is an annotated collection of all publicly available DNA sequences database	(Benson *et al.*, 2008)
RefSeq	The NCBI Reference Sequence is a non-redundant collection of richly annotated DNA, RNA, and protein sequences.	(O'Leary *et al.*, 2016)
KEGG	The Kyoto Encyclopedia of Genes and Genomes (KEGG) database integrates functional information, biological pathways, and sequence similarity in order to infer high level functions of organisms or ecosystems.	(Kanehisa and Goto, 2000)
Uniprot	Collection of Uniprot protein sequences and functional information, comprehensive and non-redundant database that contains most of the publicly available protein sequences.	(The UniProt Consortium, 2021)
IMG	The Integrated Microbial Genomes & Microbiomes is a genome browsing and annotation platform providing a comparative analysis context of assembled metagenomic data with the publicly available isolate genomes.	(Markowitz *et al.*, 2012)
eggNOGs	Evolutionary genealogy of genes: Non-supervised Orthologous is a database of orthology relationships, functional annotation, and gene evolutionary histories.	(Muller *et al.*, 2010)

The M5NR non-redundant protein database was created to enable large scale sharing of sequence data accompanied by similarity results. For this, M5NR reduces the overall resource consumption by including a significant number of the available data sources, as follow listed in the Table.

4.2. Identification and Comparison of Biosynthetic gene clusters

a. AntiSMASH tool application

To proceed with genome mining of the *Nostoc* cyanobionts species in a second part, the selected bins of cyanobacteria were submitted to the *antiSMASH* standalone version (Blin *et al.*, 2021). The antibiotics and secondary metabolite analysis shell (*antiSMASH)* is an automated genome mining pipeline and the current mainstay tool in microbial biosynthetic pathway and secondary metabolites gene cluster identification. The predictions provided by *antiSMASH* to identify BGCs candidates are driven by a set of manually curated rules which currently covers 70 different secondary metabolites gene clusters such as non-ribosomal peptides (NRPS), polyketides (PCS), terpenes, bacteriocins, cyanobactins, ectoins, nucleosides among others (Blin *et al.*, 2021). The software provides the user with desirable information such as the putative chemical structure and type of metabolite, as well as the BGC location in the genomes (Kealey *et al*., 2017). Since its initial release in 2011, six versions of *antiSMASH* have been published, and version 6 has been specially developed for improved characterization of NRPS, PKS, and RiPPs cluster types. The complete list of biosynthetic gene clusters referenced by antiSMASH is given in Annex 4.

The core logic of genome mining tools is based on the ability to detect clusters in genome data. As is the case with the majority of these tools, *antiSMASH* uses protein motifs to detect the presence of bioactive gene clusters related to secondary metabolites. The background of protein motifs is represented in profile hidden Markov models (pHMMS), these are probabilistic profiles designed to gather information about known genes by using sets of multiple sequence alignments, including variable and conserved regions while enabling an efficient search for characteristic domains of protein families (Meren Lab., 2022). Subsequently, hit genes are used as anchors from which gene clusters are extended upstream and downstream by a specified extension distance (Kautsar *et al.,* 2017). Following on from protein profile detection, the BGCs are identified according to the defined ruleset of which enzymes are required to form a metabolic pathway of a specific type. In the second stage, analyses are then carried out on the characterized biosynthetic clusters using the biochemical data to predict additional details about the produced metabolites. Figure 7 illustrates the general workflow of the latest version of *antiSMASH* and highlights the panel of prediction tools and resources integrated by the software.

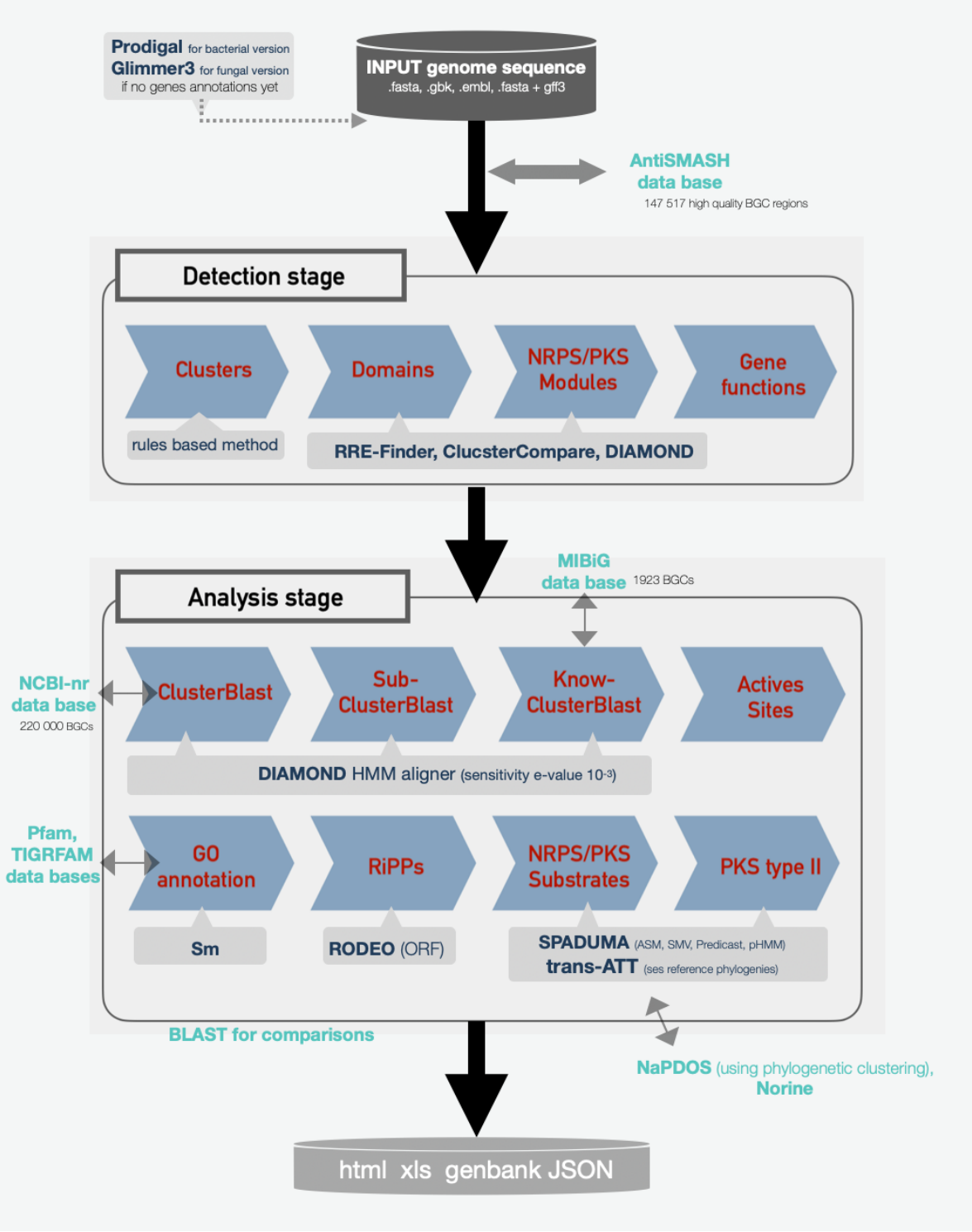


Figure 7. Schematic general workflow of antiSMASH *v 6.0* genome mining application [Figure adapted from Blin K et al., 2021]. Valid input formats are FASTA, FASTA + GFF3, GenBank, EMBL, and download via NCBI accession number. In the first step of the Detection stage potential BGCs are identified using a rules-based method. Region-specific analyses are made for NRPS/PKS domain structure, NRPS A-domain specificity prediction, PKS AT-domain specificity prediction, identification of similar clusters in antiSMASH DB or MIBiG, smCoG gene family prediction, active site classification, and GO classification. In the second Analysis stage, these BGCs are then subjected to various analyses, including predictions of function or specific BGC features, and comparative methods that can identify if similar BGCs have been described in databases.

AntiSMASH v 6.0.1 was run on FASTA files of the drafts genomes with *--prodigal-m* argument to first obtain the protein-coding gene predictions through the PROkaryotic DYnamic Programming Genefinding Algorithm -*Prodigal*, with the metagenomic *m* flag (Hyatt *et al.*, 2010). After gene prediction, the BGCs were identified with pHMMs by their signatures and four main analyses were performed: analysis of secondary metabolism protein families (smCOG), ClusterBlast analysis for purpose of BGC comparisons, NRPS/PKS domain analysis, and analysis for prediction of PKS and NRPS chemical structures (Grujic, 2021). Additionally, the next settings *--asf --pfam2go –smcog-trees --cb-general --cb-subclusters --cb-knownclusters* were specified to perform a full-featured submission which ran active site finder analysis, ran *Pfam* to Gene Ontology, generated phylogenetic trees and compared identified clusters against a database of antiSMASH-predicted clusters, against known sub-clusters, and finally against known gene clusters from the MIBiG database, respectively. In particular, the similar BGCs revealed from MIBiG annotated entries could be used to make structural predictions of the putative compounds (Grímur *et al.*, 2021).

Regardless of which analysis option is requested, various outputs are provided by *antiSMASH*. In this study firstly, the complete analyses were collected in a JavaScript Object Notation (JSON) format while a limited representation of clusters identification with detection area was given in a *regions.js* file. These representations of the regions include both the gene and protein sequences, among other data emanating from the ClusterBLAST analysis. Since these JavaScript formats are text only, JSON and JS data outputs are purposefully easy to use by other programming languages and may be sent between bioinformatic tools for further analysis. On the other hand, *antiSMASH* provides HTML files displaying research-friendly and detailed views of the respective clusters on its web interface.

Currently, the *antiSMASH* platform is the most comprehensive in the field of genome mining for BGCs. On its main page, it lists all detected clusters, as well as the most similar and known BGCs expressed in percentages. Each of the identified BGCs can be explored through the interface towards a more precise view of the linear gene arrangement, with details like protein identification or gene function (retrieved from BLAST results on the NCBI database). Other subdirectories are given for *NRPS/PKS Domain, ClusterBlast, KnownClusterBlast*, and *SubClusterBlast* analysis.

b. Improvement of antiSMASH results using Palantir

The annotation of BGCs is characterized by multidimensional and rich data structures, which can be complex to manage. Additionally, antiSMASH analysis only screens one genome at a time, so does not allow for easy comparison between samples at a large-scale level. It is therefore not suitable for a type of projects such as the present one that requires automation of the parsing, storage, and querying of such large sets of annotation data. *Palantir* (Post-processing Analysis tooLbox for ANTIsmash Reports) is dedicated software for handling and refining secondary metabolite biosynthetic gene cluster data annotated with the popular *antiSMASH* pipeline (Meunier *et al.,* 2020). This toolbox, available from metaCPAN, was used in the present large-scale genome-mining project (involving the analysis of the many *Nostoc* cyanobiont genomes) to reformat antiSMASH reports for storage in a relational database (SQL). All 53 *regions.js* files obtained from the standalone version 6 of *antiSMASH* were run on the *Nostoc* MAGs and were submitted to *Palantir's* methods using the command *export_bgc_sql_tables.pl* (which exports the BGC information into SQL tables). Moreover, this command line uses additional

Palantir functionalities through the following four default options: --*module-delineation*, --*gap-filling*, --*under-recov,* and --*undef-cleaning*. These features provided by *Palantir*, while complementary to the results performed by *antiSMASH*, aim to result in a more accurate characterization of the multimodular NRPS and PKS clusters, and therefore help to reach the goal of a more complete BGC annotation.

Subsequently, the data of the SQL tables were ported within a database and analyzed by a set of sqlite3 query commands. This SQL database was useful for both data visualization and statistic management, it lends itself to the analysis of large-scale and hierarchically organized genome mining, and in this case for identifying the biosynthetic gene clusters of Peltigera cyanobionts.

PART III: Results and discussion

5. Presentation of the results

5.1. Preprocessing of the sequencing reads

The Illumina NovaSeq 600 machine-generated paired-end reads of 151 bp with insert sizes of ~500 bp. Preprocessing of the sequencing reads resulted in a dataset of millions of reads of 131 bp length on average (Supplementary Table in Annex 6). Despite the sequencing accuracy of the Illumina technology (Phred30 before filtering ~89 %), an average of 4 % of reads per sequencing set were discarded during the control of the quality.

5.2. Assembly of the reads using metaSPADES

In the second step of the pipeline, short-read sequences were assembled using *metaSpades*. As lichens are complex microbial communities, the resulting reads for each metagenome represent eukaryotic and prokaryotic microbes, including fungi (mycobiont and endolichenic fungi) and bacteria (cyanobacterial photobiont and secondary epiphytic and endophytic bacteria). Ideally, the result of the metagenome assembly process should have been a set of genomes of all the species that are represented within the lichen thalli. In reality, metagenomic assemblies were quite fragmented (82 603 contigs per assembly on average) because of the non-uniform sequencing data involved in the complexity of such communities. To evaluate assembly performance, several assembly statistics and metrics from QUAST, i.e. total length, number of scaffolds (≥ 500 bp, ≥ 1000 bp, and ≥ 10 000 bp), N50 and N75, L50, and L75, as well as the GC content and the number of gaps per 100 kbp, were considered and reported in the table of Appendix 7. N50 is the minimum length to cover at least half of the assembly with larger contigs. It serves as a median value for assessing whether the assembly is balanced towards longer contigs (higher N50) or shorter contigs (lower N50). For obtaining full-length coding sequences, the assembly of the short reads is expected to be as contiguous as possible.

Figure 8 shows the distribution over the samples of the N50 values together with the total number of contigs. These plotted metrics made it possible to quickly identify the metagenomes most likely to respond to this contiguity such as the assemblies of samples S41, S20, and S67 characterized by high N50 and a limited number of contigs as shown by the chart, as well as the samples related to highly fragmented assembly like S34, S51, and S23 (characterized by a large number of contigs and small N50). Moreover, the visualization provided in Figure 8 highlights the strong variability of these two measures between the metagenomes.

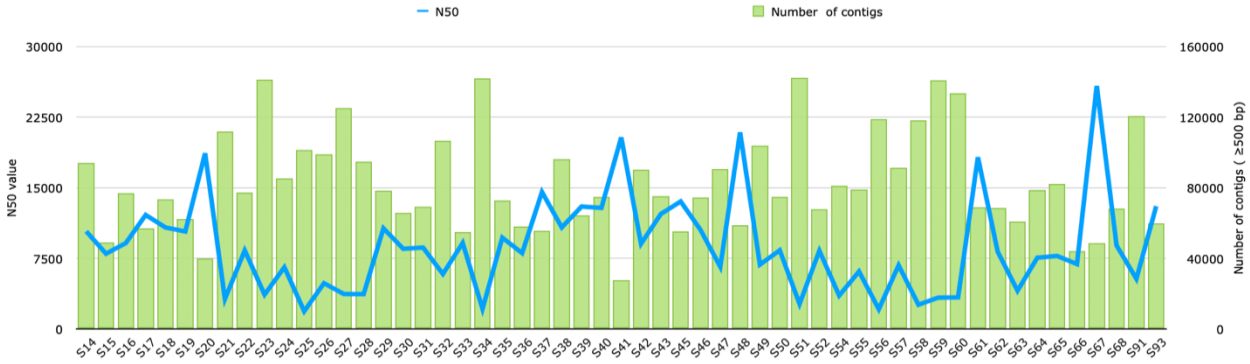


Figure 8. Distribution of the N50 values and numbers of contigs of the metagenomes. The assembly statistics reported in the chart were computed by *QUAST v 5.1*. The number of contigs and the N50 values recorded for the 56 metagenomes assembled are presented in bar plots and distribution curves respectively. The statistics were based on contigs of 500 bp and higher.

The metagenome assembly of sample S23 from the *Peltigera vitikainenii* collection was the largest (246 632 577 bp from 141 184 contigs), while the assembly of the S66 was smaller than the others (120 378724 bp from 44 116 contigs). In other collections, the largest assembly was S59 (244 730 407 bp) from *P.appalachensis*, the S61 (196 975 071 bp) from *P.asiastica*, the S68 from *P.borenquensis* (170 950 990 bp), the S64 (167 071 012 bp) from *P.mikado*, and the S14 (224207994 bp) from *P.pacifica*.

5.3. Taxonomic assignments of the reads using Kraken2

The results of the taxonomic assignment of raw reads were reported in a tab-delimited report for each of the six Peltigera species sampled and summarized in Table 5. Most of the taxonomic profiling achieved by *Kraken2* allowed the identification of only a minority of the sequences. Approximately 45% of the reads in the metagenomic samples could be classified. Yet, all of the three domains of life, i.e. Archaea, Bacteria, and Eukaryota, were represented. There were also viruses detected in all samples. Some bacterial taxa that were ubiquitous across the lichen samples of this work, namely, Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes, and Uroviricota, have also been observed in previous studies of lichen-associated bacteria (Tomislav *et al.*, 2019; Ponsero *et al.*,2021).

Table 5. Taxonomic assignment of reads recovered from metagenomic samples of six Peltigera species.

		Peltigera appalachensis	*Peltigera asiatica*	*Peltigera borinquensis*	*Peltigera mikado*	*Peltigera pacifica*	*Peltigera vitikainenii*
Unclassified		44 %	57 %	44 %	46 %	46 %	55 %
No blast hits		5 %	6 %	4 %	4 %	5 %	5 %
Archea		0.07 %	0.04 %	0.05 %	0.04 %	0.05 %	0.07 %
Bacteria		30 %	13 %	35 %	31 %	23 %	27 %
Eukaryota		21 %	21 %	17 %	17 %	14 %	13 %
Viruses		0.5 %	3.4 %	0.2 %	2 %	2 %	0.3 %
Kingdom	Fungi	0.8 %	1 %	1 %	1 %	1 %	1 %
	Metazoa	18 %	16 %	9 %	14 %	11 %	9 %
	Viridiplantae	0.8 %	1 %	4 %	0.7 %	0.8 %	0.7 %
Phylum	Actinobacteria	2 %	1 %	2 %	1 %	2 %	4 %
	Ascomycota	0.7 %	1 %	1 %	1 %	1.3 %	1 %
	Arthropoda	3 %	6 %	4 %	7 %	4 %	3 %
	Bacteroidetes	0.3 %	0.4 %	0.3 %	0.2 %	0.3 %	0.4 %
	Chordata	14 %	8 %	5 %	6 %	6 %	5 %
	Cyanobacteria	14 %	6 %	25 %	24 %	14 %	9 %
	Firmicutes	0.2 %	0.3 %	2 %	0.2 %	0.3 %	0.2 %
	Proteobacteria	12 %	4 %	5 %	5 %	5 %	7 %
	Streptophyta	0.7 %	1 %	4 %	0.7 %	0.8 %	0.7 %
	Uroviricota	0.1 %	3 %	0.1 %	2 %	0.2 %	0.2 %
	Unclassified baccteria	0.006 %	0.01 %	0.004 %	0.006 %	0.005 %	5 %
Genus	Nostocc	13 %	6 %	23 %	22 %	13 %	8 %
	Peltigera	0.3 %	0.9 %	0.9 %	0.7 %	0.9 %	0.6 %

Table: The taxonomic profiles of the raw reads set of the samples P.appalachensis_S59, P.asiatica_S61, P.borinquensis_S68, P.mikado_S64, P.pacifica_S14, and P.vitikainenii_S23, were characterized by *Kraken2* using a reference database. Sequences that do not share enough k-mers with those in this reference database were labeled as unclassified. The results are reported in percentages of reads covered by the clade rooted (rounded to one significant digit). No blast hits are the fraction of the number assigned directly to the root on the total classified reads.

The taxonomic profiles for the most abundant (>1% relative abundance) Phyla and Genera from the final assignments set are depicted in Figure 9. The majority of reads belonged to bacterial phyla among which the Cyanobacteria (up to 53%) were by far the most abundant, followed by Proteobacteria. Other phyla present across all the samples included Actinobacteria, Ascomycota, and Streptophyta, as well as phyla from the Metazoa kingdom, the Chordata, and Arthropoda. The Ascomycota were the most abundant phylum from the Fungal kingdom and the Streptophyta were the most abundant phylum from the Viridiplantae kingdom. Nevertheless, although the phylum to which the lichen-forming fungi belong, reads assigned to the Ascomycota made up a small proportion of the samples. Additional phyla were present at relatively high abundance in unique samples only. Notably, an important proportion of the reads from *P.vitikainenii* were assigned to the bacterial phylum Nitrospirae and the viral phylum Peploviricota was significantly present in the *P.pacifica* sample.

At lower taxonomic ranks, the fraction of classified sequences decreased tremendously. Microbial diversity at the genus level has been characterized by the tool with a relatively coarse resolution, typically identifying some major cyanobacterial genera essentially. A long tail of taxa present at less than 2% relative abundance was funded. Nevertheless, a proportion of the taxonomic assignments at the genus level were related to *Nostoc* lineage (total of >147M reads) and few reads appeared to belong to the Peltigera in the samples *P.pacifica* and *P.asiatica*.

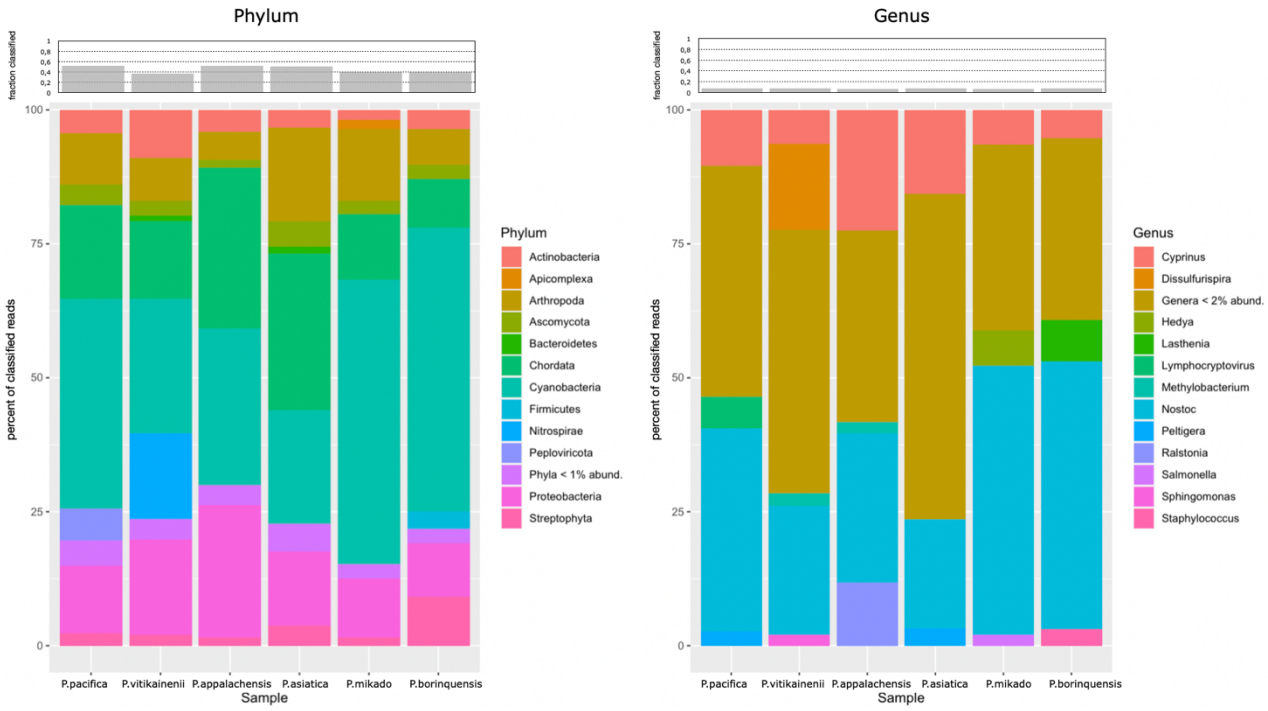


Figure 9. Relative abundances of various major lineages recovered from lichen species. Relative abundances at the phylum level are shown in the right chart and relative abundances at the genus level in the left chart. Relative abundance was calculated as the percentage of reads belonging to a particular lineage of all the classified sequences recovered from a given sample. The most abundant lineage at the phylum level was the Cyanobacteria (29%,21%,53%,53%,39% and 25%).

However, the measurements provided by metagenomic sequencing are not capable to account for the real diversity of the environment. Relative abundance is the proportion of sequences assigned to taxa, not the real abundance of the organism in the sample, and may be distorted. In addition, many potential biases in the metagenomic workflow could have led to erroneous results. For example, the presence of contaminating DNA at the sampling stage may incorrectly suggest that a microbial species is detected within the community. In the present case, the Streptophyta is a clade of plants including land plants and all green algae except the Chlorophyta, therefore seemed more to be environmental contamination rather than organisms involved in the lichen holobiont. Finally, taking into account the difficulty of the task requested and therefore its limitations, the consideration of false positives and/or false negatives was not negligible. For example, the results obtained for the assignment at the genus level of the Cyprinus seemed more likely to be false positives since it is a genus of fish.

The results obtained did not allow the identification of a large part of the presumed microbial population present in the metagenomic samples and the consideration of erroneous assignments was not negligible. The number of assignments to the leading domain of Bacteria was restricted to not exceeding 35% of the total reads. Moreover, a remarkable fraction of reads (48% to 63%) escaped the labels established by *Kraken2*, referred to as ‘unclassified’ and 'no blast hits' in Table 5, for which no matches were found in the database used. Subsequently, since this method of classifying readings has generated taxonomic profiles potentially biased by analysis limitations and the lack of information on the unrepresented species, the representation of relative abundances was also considered distorted.

5.4. Evaluation and comparison of the binning results

The sequences from the metal spades assembly were clustered into 3288 bins using *CONCOCT* and 807 bins using *metaBAT2*. The taxonomy assignment of the bins was checked using SCG taxonomy derived from the gold standard GTDB for bacterial and archaeal classification. Respectively, the bins of each of the two collections were classified according to the criteria of completeness and contamination to compare the abilities of the two binning tools. Analysis of the binning tables was performed using Rstudio software *v 1.2* (script bins_stats.R on Github). The summary of the clustering results delivered by CONCOCT and *metaBAT2* is provided in Table 6.

Table 6. Binning metrics are summarized for each of the binning tools.

	CONCOCT	Metabat2
Total bins	3288	807
Mean total bins per sample	60	15
Bins without quality assignment (completeness = 0)	2360 (~72 %)	231 (~29 %)
Bins > 90% complete	211 (~6 %)	157 (~19 %)
Bins > 50% complete	502 (~15 %)	328 (~41 %)
Bins < 10 % contaminated	757 (~23 %)	533 (~66 %)
Highly inconsistent bins (> 25 % contaminated)	128 (~14 %) *	29 (~5 %) *
High-quality bins (>90% complete and <5% contam.)	143 (~4 %)	138 (~17 %)
Medium quality bins (>70% complete and <10% contam.)	263 (~8 %)	239 (~30 %)
Bins with phylum annotations	593 (~18 %)	455 (~56 %)

Table: The abundances marked with an asterisk are calculated from the subset of bins with a quality assignment.

The number of bins built by *CONCOCT* in each of the samples was much higher than the total number of bins per sample from the clustering of *metaBAT2*. While *CONCOCT* built an average of 60 bins per assembly, the average number of bins built from metagenomes by *metaBAT2* was only 15 bins. The highest number of bins obtained for the same metagenome was 92 bins *CONCOCT* in sample S25 and 27 bins *metaBAT2* in sample S38. Despite the large number of bins generated by *CONCOCT*, there were often only a few good quality bins in the sample results. For example, 83 bins were generated by *CONCOCT* from the S66 metagenome, of which only 2 were of medium quality. Although more commonly observed in *CONCOCT* results, both *CONCOCT* and *metaBAT2* have created orphan bins, e.g. without taxonomic assignments and qualitative assignments (completion equal to zero). Moreover, despite a smaller total number of bins, the share of bins with taxonomic assignments in the *metaBAT2* collection was much higher (56% vs. 18%). Otherwise, the comprehensive annotation of 3 047 (74%) of the bins has remained unsolved.

Then, the ranking of the bins according to their estimated quality shows that *metaBAT2's* performances have globally outpaced those of *CONCOCT*, allowing in particular to obtain the greatest set of bins of high and medium quality. To closer compare the bin sets in terms of completion and contamination, the evolution curves of these data have been drawn for each of the quality estimates

provided by *CheckM* (Figure 10. A, B). The quality of eukaryotic bins assigned to the lineage of fungi was assessed by the distribution of F1 scores (Figure 10. C).

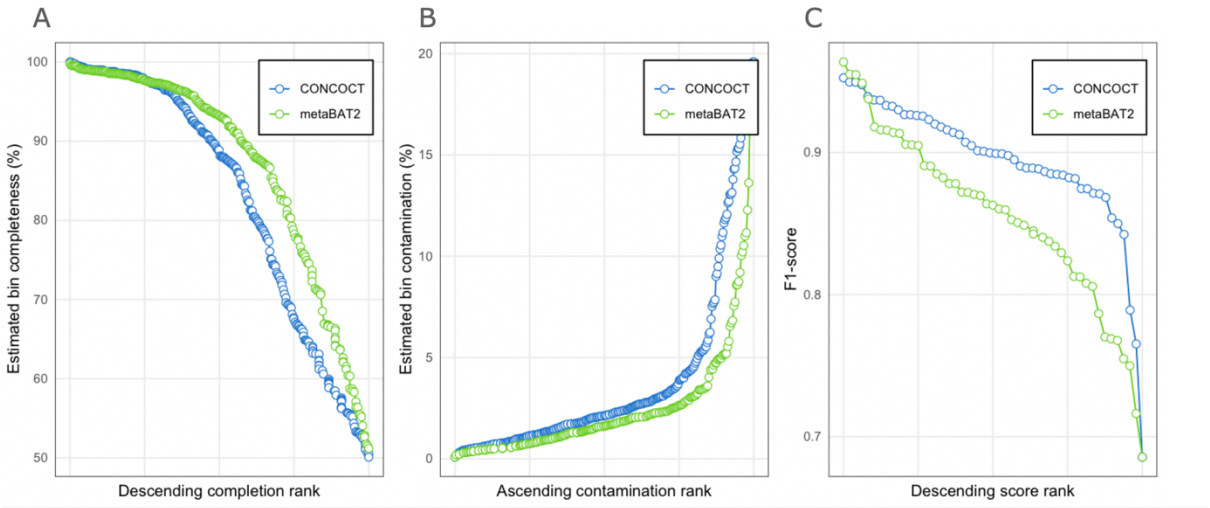


Figure 10. Comparison of binning tools for assembling draft genomes from the metagenomic dataset. Comparing bins generated by MetaBAT2 (green points) and CONCOCT (blue points) ordered by their ranked value in a normalized distribution (x-axis). Only consistent bins are showed (completeness >70% and contamination <20%). (A) Percent completeness (y-axis) of the bins ordered in decreasing order of their completeness value evaluated by CheckM. Completeness is defined as the percentage of the assigned genome represented in the bin. The upper curve indicates that more bins are binned by the tool at a higher level of completeness. (B) Percent contamination (y-axis) of the bins ordered in the increasing order of their contamination value evaluated by CheckM. Contamination of a bin is defined as the percentage of nucleotides that did not align to the assigned genome. The lower curve indicates that more nucleotides are binned by the tool at a lower level of contamination. (C) F1-scores of the bins from Fungi (Eukaryota) linages are ordered in decreasing order of their score value. F1-scores were calculated from completeness and contamination assessed by EukCC (equation p. 35). The upper curve indicates that more bins are binned at a higher level of completeness and lower level of contamination.

Comparing the quality of the bins assed by *CheckM* in Figure 10. A and B showed higher levels of completeness (green upper curve) and lower levels of contamination (green lower curve) of the bins generated by *metaBAT2*. On the other hand, *CONOCT* performed better for generating bins of higher quality when concerning Fungi lineages. Indeed, *CONCOCT* allowed the assembly of 47 completeness quality bins higher than 70%, against 36 bins in the *metaBAT2* dataset. The average completeness and e-average contamination evaluated by *EukCC* for bins associated with Fungi lineages (showed in Figure 10. C), were 83% and 2.6% for bins of the *CONCOCT* dataset, toward 77% and 2.1% for results of the *metaBAT2* dataset. In general, *metaBAT2* tended to split the fungal genome into two or three separate bins whose completeness was below the threshold imposed in the first trimming (completeness >70% and contamination <10%). Samples S16, S18, S20, S24, S37, S40, S46, S47, S49, S62, S65, and S67 are examples where no bin representative of lichen-associated fungi was found among the *metaBAT2* results in the medium quality trimming, while more than one bins fungi were present. However, no fungal bin were retained in the results of samples S25, S30, S43, S45, and S51, neither using *CONCOCT* nor using *metaBAT2*. In sample S49, no bin *metaBAT2* were binned to represent either cyanobiont or mycobiont (only 1 bin Acidobacteriota and 2 bins Alphaproteobactceria).

Across all metagenomes, the Cyanobacteria lineage was detected by both tools in 50 samples. No bin associated with Cyanobacteria was found in the *metaBAT2* results for samples S14, S49, and S50, while no assembly was allowed to represent the cyanobiont with sufficient quality in sample S36 (1 bin concoct with 23.22% completeness).

In addition to the main symbionts, a variety of lichen-associated bacterial genomes were also generated during the metagenomic binning of this study. Besides the Cyanobacteria phylogroup, the *GTDB-toolkit* has achieved the taxonomic classification of 879 binned sequences related to diverse bacterial lineages. The highest number of consistent bins (medium quality) from this bacterial colonization was observed in the samples S26 and S59 counting 14 bins generated by both binner tools. The total number of these bins in the *CONCOCT* and *metaBAT2* datasets were 158 and 149 respectively and involved 10 phyla: Proteobacteria, Verrucomicrobiota, Actinobacteriota, Acidobacteriota, Armatimonadota, Firmicutes, Bacteroidota, Planctomycetota, Eremiobacterota, Nanoarchaeota. Conversely, there was no consistent sequence classified as potential microbiome-associated bacteria in the samples S33, S41, S43, S47, and S66, but only the two major symbionts.

5.5. Manual and automated bin refining

The bins from the binning tables were then refined using the automated *DASTool* binner in addition to manual curations. Curations on the combined *CONCOCT* and *metaBAT2* libraries resulted in refined collections of medium-quality draft genomes. The resulting bins from each sample refinement were exported in Annex 7. The refined collection consisted of 261 draft genomes. From this, 53 draft genomes were related to the *Nostoc* genus, 49 to fungal lineages, and 159 draft genomes were assigned to bacterial lineages.

By the binning method comparison analysis, the majority of the selected cyanobacterial bins were generated by *metaBAT2* (46 bins *vs.* 8 bins), while the selection of eukaryote draft genomes from the Fungi lineages was dominated by bins generated by *CONCOCT* (38 bins *vs*. 9 bins). Finally, the bacterial sequences, which would be potentially representative of the specific lichen-associated microbiome, were similarly derived from both the *metaBAT2* and *CONCOCT* collections (73 and 86 bins). Average statistics are provided in Table 7 for each of these three groups.

Table 7. Summary of some statistical metrics recovered from the refine genome collection.

	Cyanobacteria	Fungi	Bacteria
Number of draft genomes :	54	47	159
Av. length :	7 624 400 bp	31 063 266 bp	4 431 391 bp
Av. Number of scaffolds :	315.24	2218.81	542.84
Av. N50 :	53 524 bp	22 754 bp	52 579 bp
Av. Completeness :	97.27 %	83.96 %	89.48 %
Av. Contamination :	0.72 %	2.38 %	4.63 %

Table: Averages were calculated on the basis of the metrics detailed in Appendix X that lists the drafts genomes selected towards the refinement of medium quality sequences. The completeness and contamination were evaluated by CheckM for the group Cyanobacteria and Bacteria, while evaluated by EukCC software for the draft genomes of Fungi group.

For each of the sample collections, the draft genome assigned to the *Nostoc* lineage was assumed to be the cyanobiont partner of the symbiosis, and the draft genome assigned to the Fungi lineage to represent the lichenicolous mycobiont. Measures of a quantitative assessment of genome completeness and complementary intuitive metrics to describe these genomes were provided by *BUSCO* (The Benchmarking Universal Single-Copy Orthologs) and were plotted in Figure 11. The total

length of the draft genomes was higher than 7.5 Mbp and 31.06 Mbp, the number of scaffolds in the assemblies was 269 and 2219 on average, and the N50 were 54.5 kbp and 22.8 kbp on average for the cyanobacterial and the fungal genomes respectively.

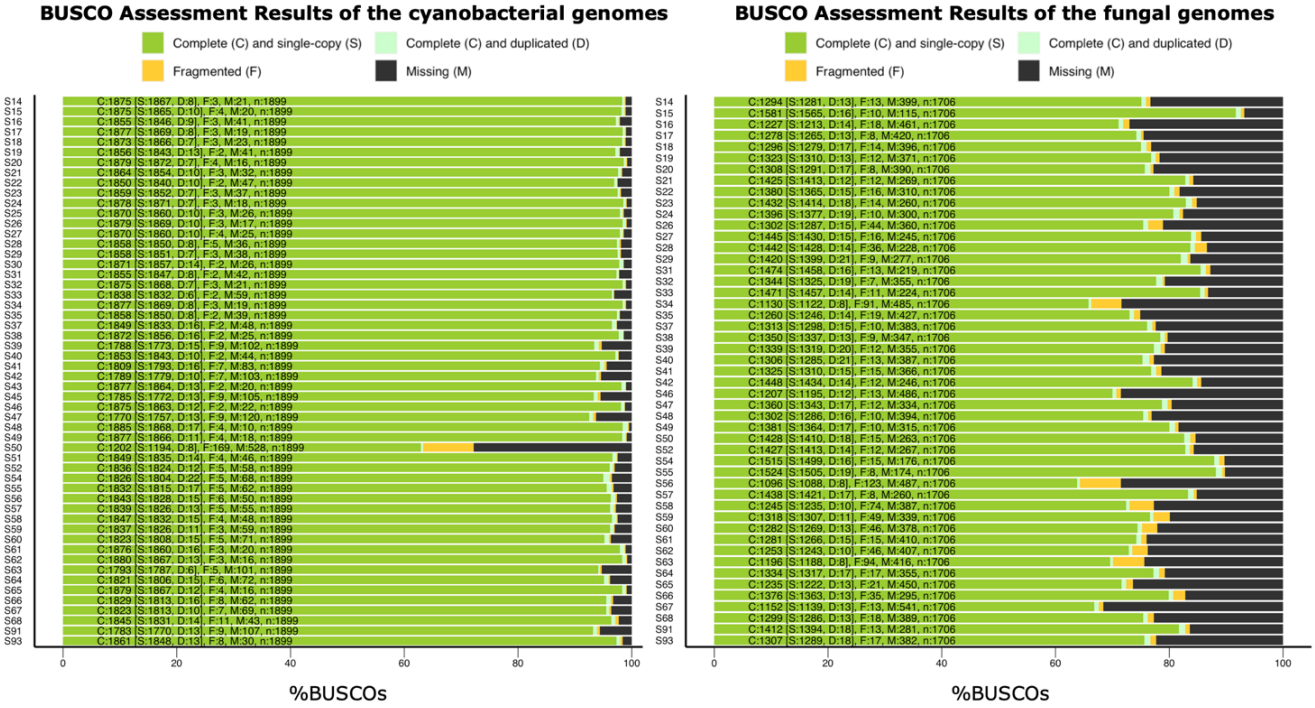


Figure 11. BUSCO assessment of the cyanobiont and mycobiont draft genomes from the refined collection. C:complete [D:duplicated], F:fragmented, M:missing, n:number of genes used. The recovered genes are classified as 'complete' when their lengths are within two standard deviations of the BUSCO group's mean length. 'Complete' genes found with more than one copy are classified as 'duplicated'. Genes only partially recovered are classified as 'fragmented', and genes not recovered are classified as 'missing'. Finally, the 'number of genes used' indicates the resolution and hence is informative of the confidence of these assessments.

The draft genome assemblies of the fungal partners appeared less consistent than the cyanobacterial genomes. Excluding the genome from the sample S15 which was uniquely completed (93%), the maximum percentage completeness of the fungal genomes set was 88% while the percentage completeness of the cyanobacterial genomes was 97% on average. The assemblies of fungal draft genomes were more fragmented. Furthermore, the proportion of the fungal genomes not recovered by the *BUSCO* gene models (classified as 'missing' and colored in grey in Figure 11) was remarkably high (7% to 29%), which illustrates the poor representation of lichenicolous fungi in current databases. For these reasons, fungal draft genomes were discarded for further in-depth analyses, leaving this study to focus only on high-quality genomes, thus on the cyanobacterial metagenome-assembled genomes (MAGs). Finally, the cyanobiont genome from the sample S50 also contained a significant number of genes classified as missing (28%) but it was also very fragmented (9%) in comparison with other genomes of the cyanobacterial set. This is why this draft genome from sample S50 was removed from the set of cyanobiont genomes with the intention of downstream analysis.

5.6. Selection of the metagenomic-assembled genomes of the Nostoc cyanobionts

A final selection consisting of 53 MAGs was kept for the remainder of the study. This collection of MAGs, listed in Table 8, recruited 404 716 742 nucleotides, which represented only 1.87% of all nucleotides stored in the assembly collection. The final collection consisted of high-quality draft genomes (completeness > 90% and contamination < 5 %) taxonomically classified to the *Nostoc* genus. Twenty-three of these MAGs were assigned to species-level using *FastANI* (Jain *et al.*, 2018) within the *GTDB-toolkit* with ANI > 95 % and an alignment fraction >65%. The cyanobionts related to the species *Peltigera Pacifica* and *Peltigera vitikainenii* were very similar to the *Nostoc strain sp002949735* (GTDB reference GCF_002949735.1) referred in the NCBI to the organism *Nostoc sp*.' *Peltigera* membranacea cyanobiont' N6, collected in Iceland (Gagunashvili and Andrésson, 2018). Whereas, the harvest territory of the specimens considered in the present work extends throughout the northern hemisphere. Based on *GTDB-Tk* classification, the thirteen remainders of the MAGs represent potential new species that do not have representatives in known databases. These MAGs represent a dataset that can be interrogated using any of the available databases, or classified in a phylogenetic tree through comparative genomics. Indeed, the mining of these genomes for bioactive secondary metabolites was the focus of this study.

Table 8. Genome statistics of the selected MAGs belonging to the genus of *Nostoc*

Sample libraries	Binning software	Bin	Genome size (Mbp)	Number of scaffolds	N50	GC content	Median coverage	CheckM lineage	CheckM contamination	CheckM completeness	BUSCO completeness	DASTool completeness	GTDB lineage
Peltigera appalachensis :													
S39	METABAT	8	8.191194	431	29719	41.90 %	156.13	p__Cyanobacteria	0.41 %	94.89 %	94.15 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S40	METABAT	18	7.299480	135	79090	41.94 %	149.85	p__Cyanobacteria	0.44 %	98.00 %	97.57 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S41	METABAT	8	7.618213	374	29897	41.93 %	110.97	p__Cyanobacteria	0.41 %	94.22 %	95.26 %	94.12 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S42	METABAT	5	7.475994	397	27916	41.95 %	162.5	p__Cyanobacteria	0.48 %	93.56 %	94.2 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S43	METABAT	10	8.524251	226	67551	41.62 %	61.91	p__Cyanobacteria	0.44 %	98.78 %	98.84 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S45	METABAT	5	7.988487	390	29870	41.85 %	164.91	p__Cyanobacteria	0.49 %	93.22 %	93.99 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S46	METABAT	3	8.493280	219	67584	41.62 %	125.49	p__Cyanobacteria	0.44 %	98.33 %	98.73 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S47	METABAT	6	7.275875	410	25925	42.05 %	92.43	p__Cyanobacteria	0.37 %	92.66 %	93.2 %	94.12 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S48	METABAT	5	8.164301	136	106111	41.44 %	94.91	p__Cyanobacteria	0.52 %	99.78 %	99.26 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S49	CONCOCT	25	8.973099	321	57012	41.41 %	41.35	p__Cyanobacteria	0.44 %	99.22 %	98.84 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S51	METABAT	10	7.928416	304	50267	41.69 %	108.98	p__Cyanobacteria	0.33 %	98.88 %	97.36 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S52	METABAT	7	6.937367	169	64940	41.87 %	246.9	p__Cyanobacteria	0.56 %	97.33 %	96.68 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S54	METABAT	17	8.520291	332	44111	41.62 %	77.78	p__Cyanobacteria	2.19 %	96.88 %	96.15 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S55	METABAT	6	7.047296	193	57966	41.83 %	122.37	p__Cyanobacteria	1.22 %	97.11 %	96.47 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S56	CONCOCT	30	8.640538	419	35915	41.77 %	209.66	p__Cyanobacteria	1.00 %	97.33 %	97.05 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S57	METABAT	8	7.889065	287	47851	41.71 %	132.77	p__Cyanobacteria	1.30 %	96.66 %	96.84 %	94.12 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S58	METABAT	4	7.410359	237	47831	41.80 %	110.88	p__Cyanobacteria	0.56 %	97.22 %	97.26 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S59	METABAT	1	7.356769	247	46023	41.89 %	105.46	p__Cyanobacteria	0.63 %	97.33 %	96.73 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S60	METABAT	1	7.143502	214	64475	41.83 %	131.06	p__Cyanobacteria	0.56 %	96.77 %	95.99 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S91	METABAT	21	7.613700	412	26247	41.99 %	49.91	p__Cyanobacteria	0.48 %	93.77 %	93.89 %	94.12 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S93	CONCOCT	26	8.970127	463	31444	41.95 %	81.53	p__Cyanobacteria	0.56 %	98.89 %	97.99 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
Peltigera asiatica :													
S61	METABAT	14	7.618852	233	59126	41.50 %	39.56	p__Cyanobacteria	0.44 %	98.88 %	98.78 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S62	METABAT	13	7.443291	131	113067	41.59 %	99.18	p__Cyanobacteria	0.52 %	98.88 %	98.99 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S63	METABAT	1	7.627861	405	28036	42.00 %	199.08	p__Cyanobacteria	0.22 %	95.11 %	94.41 %	94.12 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
Peltigera borinquensis :													
S67	METABAT	8	7.928955	437	26155	41.99 %	83.56	p__Cyanobacteria	0.52 %	96.67 %	95.99 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S68	CONCOCT	10	8.805062	621	23270	41.99 %	155.05	p__Cyanobacteria	0.63 %	98.56 %	97.15 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
Peltigera mikado :													
S64	METABAT	6	8.033583	372	32342	41.93 %	177.76	p__Cyanobacteria	0.78 %	96.00 %	95.89 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S65	METABAT	1	8.087613	204	91693	41.54 %	163.61	p__Cyanobacteria	0.52 %	98.89 %	98.94 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S66	METABAT	4	7.602722	372	30145	42.01 %	75.82	p__Cyanobacteria	1.33 %	96.22 %	96.31 %	96.08 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
Peltigera pacifica :													
S14	CONCOCT	16	7.810121	250	56928	41.57 %	113.4	p__Cyanobacteria	0.37 %	98.78 %	98.73 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S15	CONCOCT	68	8.068875	363	37058	41.52 %	43.31	p__Cyanobacteria	0.22 %	98.78 %	98.73 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S16	METABAT	4	7.799941	256	54198	41.51 %	94.86	p__Cyanobacteria	0.41 %	98.56 %	97.68 %	68.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
Peltigera vitikainenii :													
S17	METABAT	4	6.779284	148	74729	41.53 %	144.45	p__Cyanobacteria	0.52 %	98.78 %	98.84 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S18	METABAT	17	6.859150	153	70820	41.57 %	65.13	p__Cyanobacteria	0.52 %	98.56 %	98.63 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S19	METABAT	12	6.860845	159	68093	41.59 %	140.4	p__Cyanobacteria	1.89 %	97.89 %	97.73 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S20	METABAT	11	6.988293	153	75659	41.60 %	161.83	p__Cyanobacteria	0.30 %	99.00 %	98.94 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S21	METABAT	6	7.993711	273	45256	41.54 %	61.88	p__Cyanobacteria	0.52 %	97.89 %	98.15 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S22	METABAT	6	8.234010	313	42825	41.49 %	79.54	p__Cyanobacteria	0.67 %	97.89 %	97.41 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S23	METABAT	12	6.827622	148	69602	41.60 %	68.96	p__Cyanobacteria	0.52 %	97.44 %	97.89 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S24	METABAT	4	7.049992	167	65017	41.58 %	127.07	p__Cyanobacteria	0.52 %	99.00 %	98.89 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S25	METABAT	5	8.168787	302	44872	41.52 %	108.47	p__Cyanobacteria	0.74 %	98.78 %	98.47 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S26	CONCOCT	49	8.867345	332	44913	41.54 %	146.85	p__Cyanobacteria	0.77 %	99.00 %	98.94 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S27	METABAT	4	7.292649	179	60413	41.82 %	93.77	p__Cyanobacteria	0.22 %	99.56 %	98.47 %	96.06 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S28	METABAT	7	7.315872	264	45835	41.58 %	295.04	p__Cyanobacteria	0.52 %	98.56 %	97.84 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S29	METABAT	12	6.790194	159	77166	41.63 %	208.55	p__Cyanobacteria	0.41 %	98.33 %	97.84 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S30	METABAT	7	7.074277	196	59630	41.50 %	153.5	p__Cyanobacteria	2.07 %	98.78 %	98.52 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S31	METABAT	13	8.091091	288	56053	41.64 %	135.87	p__Cyanobacteria	0.52 %	98.11 %	97.68 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S32	METABAT	13	6.777123	197	49945	41.61 %	171.89	p__Cyanobacteria	0.22 %	98.78 %	98.73 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S33	METABAT	3	6.780332	201	50525	41.66 %	296.5	p__Cyanobacteria	0.30 %	97.44 %	96.78 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S34	METABAT	2	6.725415	153	70018	41.54 %	218.25	p__Cyanobacteria	0.30 %	98.78 %	98.84 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S35	METABAT	6	6.868135	184	69455	41.61 %	123.41	p__Cyanobacteria	0.41 %	98.33 %	97.84 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S37	METABAT	8	6.966227	161	77827	41.51 %	184.15	p__Cyanobacteria	2.07 %	97.00 %	97.36 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S38	METABAT	17	7.117908	167	79095	41.53 %	194.18	p__Cyanobacteria	2.07 %	98.56 %	98.57 %	98.04 %	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735

5.7. Screening MAGs for biosynthetic gene clusters

Mining the metagenomic-assembled genomes from the *Nostoc* genus resulted in the identification of a high number of putative biosynthetic gene clusters in all specimens. The number and abundance of cluster families detected in each MAG are tabulated in Figure 12. A. About 20 clusters per genome were annotated by the bacterial version of *antiSMASH*, ranging from 14 up to as many as 32 gene clusters. A total of 1054 BGCs were detected in the 53 MAGs. Of these, 61% of the BGCs had similarities to known clusters in the MIBG database. Of the 653 putative known compounds, 193 had 100% similarity with the referred secondary metabolites. The compounds synthesized by the remaining 401 clusters were unknown. These predicted, but orphan BGCs were named arbitrarily. Several large contigs had more than one cluster or contained hybrid clusters, and therefore more than one cluster type was counted in a single region by the *antiSMASH* algorithm.

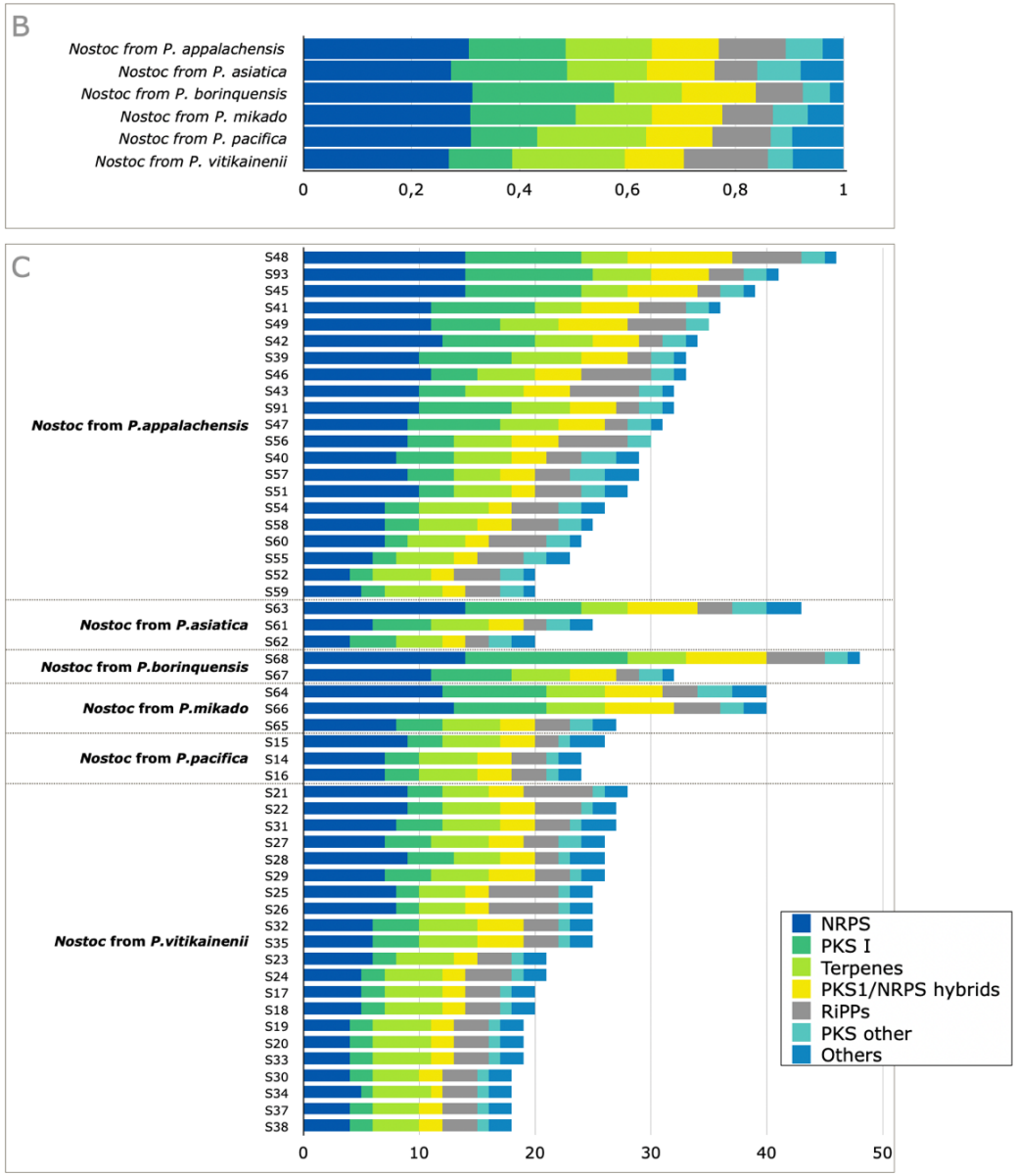
The most abundant family of secondary metabolite enzymes identified in all assembled genomes was nonribosomal peptide synthases (NRPS) followed by type 1 polyketides synthases (PKS 1). Other abundant biosynthetic gene families in the six collections of lichen, were terpenes, PKS/NRPS hybrids, RiPPs, and PKS other types (Figure 12. B). The number of biosynthetic gene clusters is illustrated in Figure 12. C was relatively similar in all the genomes except for the two leading enzyme families for which considerable divergences could be observed between and within collections of lichen species. For example, both the highest (14 clusters) and the smallest (4 clusters) number of NRPS clusters were indeed found in assembled genomes belonging to the collection of *P.appalachensis*, while the number of clusters of this NRPS family was under the average (~8 NRPS clusters on average) in the specimens retained from the *P.vitikainenii* species. Similarly, the richness of the predicted PKS-1 cluster was particularly lowest in this collection (~3 PKS-1 clusters) whereas this richness was highly variable (from 2 up to 11 PKS-1 clusters) in the collection of *P.appalachensis.*

Four of the fifty-three MAGs harbored a particularly high biosynthetic richness compared to other genomes with more than 40 gene clusters. Two of them belonged to the collection of *P.appalachensis* species (*Nostoc* from samples S48 and S93), as well as two supplementary samples owning from *P.asiatica* (S68) and *P.borenquensis* (S63) species respectively. On the contrary, MAGs of *P.pacifica* and *P.vitikainenii* cyanobiont-associated species comprised far fewer cluster types (< 30 BGCs). The biosynthetic gene clusters of these samples were further detailed in Tables 9-11.

Figure 12. Overview of the gene cluster families.

A

Sample	NRPS	PKS 1	Terpene	PKS/NRPS hybrids	RiPPs	PKS other	Others	Total :
Nostoc from P. appalachensis :								
S39	10	8	6	4	2	2	1	33
S40	8	5	5	3	3	3	2	29
S41	11	9	4	5	4	2	1	36
S42	12	8	5	4	2	2	1	34
S43	10	4	5	4	6	2	1	32
S45	14	10	4	6	2	2	1	39
S46	11	4	5	4	6	2	1	33
S47	9	8	5	4	2	2	1	31
S48	14	10	4	9	6	2	1	46
S49	11	6	5	6	5	2	0	35
S51	10	3	5	2	4	2	2	28
S52	4	2	5	2	4	2	1	20
S54	7	3	6	2	4	2	2	26
S55	6	2	5	2	4	2	2	23
S56	9	4	5	4	6	2	0	30
S57	9	4	4	3	3	3	3	29
S58	7	3	5	3	4	2	1	25
S59	5	2	5	2	3	2	1	20
S60	7	2	5	2	5	2	1	24
S91	10	8	5	4	2	2	1	32
S93	14	11	5	5	3	2	1	41
Total :	198	116	103	80	80	44	25	646
Nostoc from P. asiatica :								
S61	6	5	5	3	2	2	2	25
S62	4	4	4	2	2	2	2	20
S63	14	10	4	6	3	3	3	43
Total :	24	19	13	11	7	7	7	88
Nostoc from P. borinquensis :								
S67	11	7	5	4	2	2	1	32
S68	14	14	5	7	5	2	1	48
Total :	25	21	10	11	7	4	2	80
Nostoc from P.mikado								
S64	12	9	5	5	3	3	3	40
S65	8	4	5	3	3	2	2	27
S66	13	8	5	6	4	2	2	40
Total :	33	21	15	14	10	7	7	107
Nostoc from P. pacifica :								
S14	7	3	5	3	3	1	2	24
S15	9	3	5	3	2	1	3	26
S16	7	3	5	3	3	1	2	24
Total :	23	9	15	9	8	3	7	74
Nostoc from P. vitikainenii :								
S17	5	2	5	2	3	1	2	20
S18	5	2	5	2	3	1	2	20
S19	4	2	5	2	3	1	2	19
S20	4	2	5	2	3	1	2	19
S21	9	3	4	3	6	1	2	28
S22	9	3	5	3	4	1	2	27
S23	6	2	5	2	3	1	2	21
S24	5	2	5	2	4	1	2	21
S25	8	2	4	2	6	1	2	25
S26	8	2	4	2	6	1	2	25
S27	7	4	5	3	3	2	2	26
S28	9	4	4	3	2	1	3	26
S29	7	4	5	4	3	1	2	26
S30	4	2	4	2	3	1	2	18
S31	8	4	5	3	3	1	3	27
S32	6	4	5	4	3	1	2	25
S33	4	2	5	2	3	1	2	19
S34	5	1	5	1	3	1	2	18
S35	6	4	5	4	3	1	2	25
S37	4	2	4	2	3	1	2	18
S38	4	2	4	2	3	1	2	18
Total :	127	55	98	52	73	22	44	471



Cluster types identified by antiSMASH were classified into GCFs (gene cluster family) according to the same rules as applied by BiG-SCAPE (see Appendix 9).

Although having the largest number of BGCs, sample S68 had the least reliable *antiSMASH* results. Indeed, the majority of the clusters assigned to it were located in the scaffold with extended on both sides to the extremities of contigs regions. This observation can easily be highlighted by taking into account the lengths (bp) of BGCs obtained by comparison with the coordinates of the region on the contig. Thus, of the many BGCs obtained in the MAGs S68, most of them had cluster length values equal to the length of the region considered (Table 12 columns «Coordinates» and «Cluster Length»). The MAGs S68 possessed the number of scaffolds associated with both a small N50 value (Table 8), which implied a fragmented metagenome-reconstructed genome. Therefore, the assignment of genomic regions to biosynthetic clusters was truncated by the lack of gene contiguity within this MAG. Conversely, the genomic sequence of the S48 sample presented these metrics with remarkable quality, while having benefited from a broad attribution of biosynthetic potential. Thus, the MAG S48 was a good analytical model.

An overview of sampling metabolites clusters provided by these samples has demonstrated the impressive number of gene clusters related to major metabolic pathways of nonribosomal peptides synthases (NRPS) and poliketides synthases (PKS). NRPS and T1PKS are multi-domain enzymes that share a typical modular organization responsible to produce complex poliketides and peptides characterized by a vast range of biological activities and structural diversity. Moreover, some molecules are produced by the combination of these two biosynthetic pathways (NRPS/PKS-hybrids), such as nodularin associated with the cluster type 'T1PKS, NRPS, NRPS-like' in the sample S48. More than ten enzymatic domains were arranged in these BGCs which spanned 72 kbp. The modular and highly repetitive architecture of such NRPS and PKS clusters makes them an excellent candidate for the bioprospection of novel compounds. Yet, when considered together, these biosynthetic processes account for the majority of all cyanobacterial secondary metabolites.

Furthermore, four well-known clusters were ubiquitous in all the *Nostoc* strains, namely, the PKS cluster of heterocyst glycolipids (BGC0000869), the terpene cluster associated with the production of geosmin (BGC0000661), the NRPS cluster of anabaenopeptin NZ857-nostamide A (BGC0001479), and the NRPS-T1PKS hybrid cluster associated to nostopeptolide A2 (BGC0001028). These related natural products have been extensively studied previously in Cyanobacteria, and links with known biosynthetic clusters in the model-organism *Nostoc punctiforme* have been already assigned experimentally (Campbell *et al.,* 1997; Giglio *et al.*, 2008; Rouhiainen *et al.*, 2010; Hunsucker *et al.*, 2004).

Glycolipids synthases are involved in atmospheric nitrogen fixation when recruited by heterocysts (cells specialized in N_2 fixation). For enabling cyanobacteria to grow in the absence of dissolved nitrogenous compounds, heterocyst function requires a specific cell envelope formed by these glycolipids and other long-chain polysaccharides (Shvarev *et al.*, 2018).

Geosmin is an organic volatile compound produced by a wide range of microorganisms in the terrestrial environment. While the exact function of this compound in Cyanobacteria is still unknown, geosmin has been extensively studied due to its responsibility for earthy tastes and odors in potable water supplies (Giglio *et al.*, 2008).

The individual NRPS enzyme complex of anabaenopeptin and nostamide A can synthesize these metabolites simultaneously through the same biosynthetic pathway. Nostamide is a shorter structural homolog of anabaenopeptins which are widespread in cyanobacteria and show an impressive diversity in bioactivity (Shishido *et al.*, 2017). For example, these cyclic peptides involved in intra- and interspecific interactions of *Nostoc sp*, have demonstrated inhibitory activity towards phosphatases and proteases, which could be related to their toxicity and adaptiveness (Monteiro *et al.*, 2021)

Additionally, the secondary metabolite microviridin K and its corresponding BGC (BGC0000594) were also widely found in the dataset. Like many other cyanotoxins, microviridins are synthesized from the ribosomal peptide pathway (RiPP). The members of the RiPP family form a group structurally very diverse and complex of products. Many of them possess numerous biological features and have been considered among the most promising peptides found in cyanobacteria to serve as valuable pharmaceutical leads (Cavalcante do Amaral *et al.*, 2020). These oligopeptides are potent inhibitors of eukaryotic protein proteases and are suspected to act as tumor promoters (Kaasalainen *et al.* 2012). In this study, several orphan BGCs were assigned to RiPPs gene cluster types like the cyanobactin, LAP, RRE-containing, and protein types, as well as various lanthipeptides classes. Therefore, such potential untapped BGCs could be envisaged as candidate RiPPs clusters well suited for genome mining for pharmaceutical interest.

Further analysis found that several clusters were also present with similarity scores <50%, for examples: malyngamide, crocacin, nodularin, PcpA, scytophycin, myxothiazol, pelgipeptin, paenilamicin, thaxteramide, bartoloside, etc . These remaining clusters henceforth suggested in turn astounding natural product diversity.

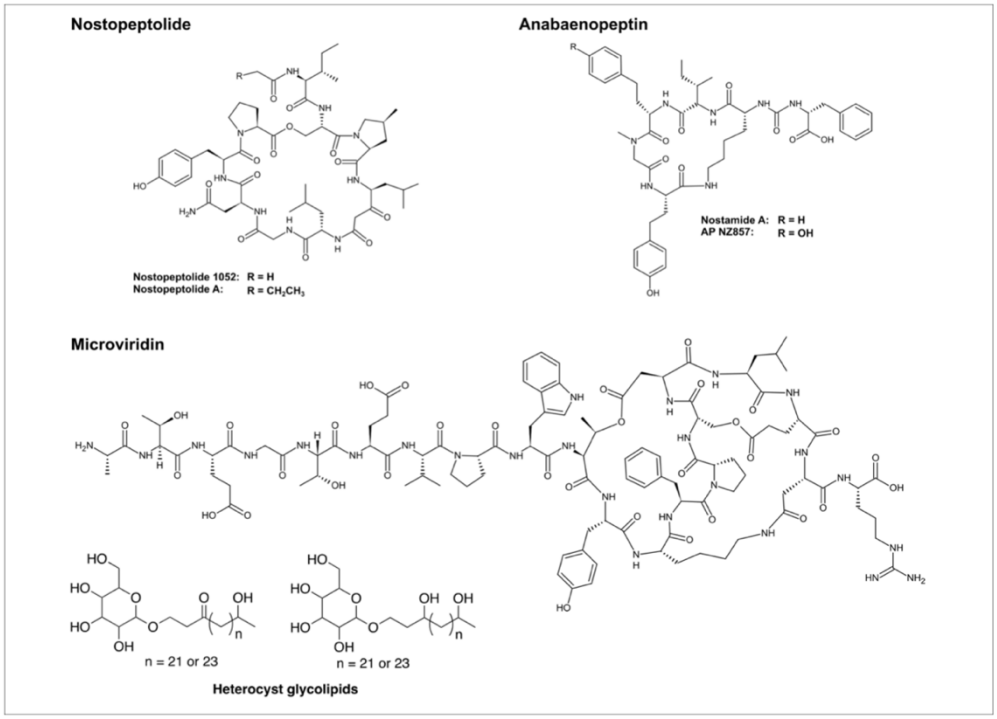


Figure 13. Natural products derived from putative BGCs (>50 % similarity with reported ones) are ubiquitous in the *Nostoc* MAGs collection.

Table 9. Putative biosynthetic gene clusters attributed using antiSMASH from *Nostoc*-associated MAG-S48

Contig id	Contig length	Contig coverage	Region	Cluster type (AntiSMASH)	Coordinates	Cluster length	Most similar known cluster	MIBiG cluster type	MIBiG accession	% Similarity
NODE_2	278329	59.0527	Region 2.1	NRPS,T1PKS,NRPS-like	85225 - 141391	56167	malyngamide	NRP + Polyketide	BGC0001971	8 %
			Region 2.2	NRPS-like,transAT-PKS-like,T1PKS	160471 - 233048	72578	crocanin	NRP + Polyketide	BGC0001971	42 %
			Region 2.3	NRPS,T1PKS	23511 - 278329	254819	nostopeptolide	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	75 %
NODE_4	264475	60.2665	Region 3.1	T1PKS,NRPS,NRPS-like	192026 - 264475	72450	nodularin	NRP + Polyketide	BGC0001705	88 %
NODE_8	216635	58.6664	Region 6.1	NRPS	21321 - 66174	44854	anabaenopeptin NZ857,nostamide A	NRP	BGC0001479	100 %
NODE_16	174973	60.1755	Region 9.1	phosphonate	1 - 26035	26035	NA	NA	-	-
			Region 9.2	RRE-containing,spliceotide	63307 - 85239	21933	PcpA	RiPP	BGC0001733	40 %
NODE_20	161156	60.1755	Region 12.1	terpene	1 - 20406	20406	NA	NA	-	-
NODE_44	118927	62.59	Region 17.1	lanthipeptide-class-v	67507 - 109797	42291	NA	NA	-	-
NODE_56	108232	61.4456	Region 24.1	NRPS	9493 - 55465	45973	NA	NA	-	-
NODE_60	106111	58.5578	Region 25.1	lanthipeptide-class-ii	256 - 25883	25628	NA	NA	-	-
NODE_61	105825	59.3867	Region 26.1	terpene	16634 - 38895	22262	geosmin	terpene	BGC0000661	100 %
NODE_72	101930	58.1378	Region 27.1	hglE-KS,T1PKS	63588 - 101930	38343	heterocyst glycolpids	Other	BGC0000869	100 %
NODE_134	82870	59.2327	Region 34.1	NRPS,T1PKS	1 - 71349	71349	aeruginoside 126B/aeruginoside 126A	NRP:Glycopeptide + Polyketide:Other + Saccharide:Hybrid/tailoring	BGC0000297	41 %
NODE_165	77600	61.0718	Region 37.1	NRPS,T1PKS,NRPS-like	1 - 41436	41436	hapalosin	NRP:Cyclic depsipeptide + Polyketide:Modular type I	BGC0001467	40 %
NODE_199	72089	59.3922	Region 42.1	terpene	39195 - 61111	21917	NA	NA	-	-
NODE_204	70627	60.378	Region 43.1	NRPS-like	11357 - 54434	43078	NA	NA	-	-
NODE_265	63295	60.2912	Region 49.1	T1PKS,NRPS-like	1 - 61819	61819	aranazole A/aranazole B/ aranazole C/aranazole D	NRP + Polyketide	BGC0001884	
NODE_414	52665	60.2691	Region 54.1	terpene	1 - 13656	13656	NA	NA	-	-
NODE_617	43623	62.3001	Region 63.1	NRPS,T1PKS	6976 - 43623	36648	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	87 %
NODE_762	38529	57.4592	Region 69.1	microviridin	26942 - 38529	11588	NA	NA	-	-
NODE_779	37983	58.2856	Region 70.1	T1PKS,NRPS	1 - 37983	37983	butyrolactol A	Polyketide	BGC0001537	33 %
NODE_1166	29480	58.8035	Region 79.1	NRPS	1 - 2948	2948	cyanopeptin	NRP	BGC0000331	75 %
NODE_1213	28540	59.9041	Region 81.1	thiopeptide,thioamitides	1 - 2854	2854	NA	NA	-	-
NODE_1349	26190	59.2713	Region 83.1	lanthipeptide-class-v	1 - 2619	2619	NA	NA	-	-
NODE_2003	17877	60.7409	Region 99.1	T1PKS,NRPS,NRPS-like	1 - 17877	17877	hapalosin	NRP:Cyclic depsipeptide + Polyketide:Modular type I	BGC0001467	60 %

Table 10. Putative biosynthetic gene clusters attributed using antiSMASH from Nostoc MAG-S66

Contig	contig length	contig coverage	Région	Cluster type (AntiSMASH)	Coordinates	length	Most similar known cluster	MIBiG cluster type	MIBiG Accession	% Similarity
NODE_3	99916	119.47	Region 1.1	transAT-PKS,NRPS-like	1 - 72974	72974	scytophycin	Polyketide	BGC0001772	27 %
NODE_5	93061	119412	Region 3.1	terpene	12349 - 33314	20966	NA	NA		-
NODE_7	82598	148877	Region 5.1	microviridin	19107 - 39256	20150	microviridin K	RiPP	BGC0000594	50 %
NODE_8	76336	119425	Region 6.1	NRPS,T1PKS,NRPS-like	1 - 40809	40809	malyngamide I	NRP + Polyketide	BGC0001971	8 %
NODE_9	73808	120349	Region 7.1	T1PKS,NRPS-like,transAT-PKS-like	1 - 61809	61809	crocacin	NRP + Polyketide	BGC0000974	38 %
NODE_11	69589	119525	Region 9.1	NRPS	43303 - 69589	26287	anabaenopeptin NZ857/ nostamide A	NRP	BGC0001479	100 %
NODE_17	59202	119.04	Region 15.1	NRPS,T1PKS	12748 - 59202	46455	paenilamicin	NRP + Polyketide	BGC0001033	28 %
NODE_18	28910	122.64	Region 16.1	NRPS,T1PKS	1 - 5891	5891	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	50 %
NODE_29	50756	121347	Region 25.1	terpene	30712 - 50756	20045	NA	NA		-
NODE_36	47436	121967	Region 30.1	lanthipeptide-class-v	18829 - 47436	28608	NA	NA		-
NODE_41	46633	117683	Region 34.1	NRPS	5426 - 46633	41208	aeruginoside 126B/aeruginoside 126A	NRP:Glycopeptide + Polyketide:Other + Saccharide:Hybrid/tailoring	BGC0000297	11 %
NODE_44	45279	121.94	Region 36.1	T1PKS,NRPS,NRPS-like	1 - 45061	45061	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	62 %
NODE_45	44489	121272	Region 37.1	NRPS-like,T1PKS,lanthipeptide-class-ii	1 - 44489	44489	jamaicamide A/jamaicamide B/ jamaicamide C	NRP + Polyketide	BGC0001001	23 %
NODE_60	38699	115004	Region 48.1	NRPS,T1PKS	1 - 35067	35067	myxothiazol	NRP + Polyketide:Modular type I	BGC0001024	28 %
NODE_70	36100	116952	Region 56.1	T1PKS	1 - 361	361	scytophycin	Polyketide	BGC0001772	55 %
NODE_81	33686	123492	Region 63.1	phosphonate	1 - 28725	28725	NA	NA		-
NODE_95	30684	119.67	Region 72.1	T1PKS	1 - 30684	30684	bartoloside 2/bartoloside 3/ bartoloside 4	Other	BGC0001525	18 %
NODE_117	27156	150621	Region 86.1	T1PKS,NRPS	1 - 27156	27156	nostophycin	NRP + Polyketide	BGC0001029	27 %
NODE_128	25729	120027	Region 91.1	hglE-KS,T1PKS	1 - 25729	25729	heterocyst glycolipids	Other	BGC0000869	85 %
NODE_129	25669	121512	Region 92.1	terpene	1 - 16329	16329	geosmin	terpene	BGC0000661	100 %
NODE_233	19548	116389	Region 123.1	terpene	1 - 19548	19548	NA	NA		-
NODE_236	19475	136903	Region 125.1	NRPS	1 - 19475	19475	NA	NA		-
NODE_428	15749	118645	Region 170.1	NRPS	1 - 15749	15749	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	37 %
NODE_792	12758	116051	Region 204.1	NRPS	1 - 12758	12758	anabaenopeptin NZ857/ nostamide A	NRP	BGC0001479	100 %
NODE_3739	6833	116.57	Region 319.1	reodx-cofactor	1 - 6833	6833	lankacidin C	NRP + Ploketide	BGC0001100	13 %

Table 11. Putative biosynthetic gene clusters attributed using antiSMASH from Nostoc MAG-S68

Contig	contig length	contig coverage	Région	Cluster type (AntiSMASH)	Coordinates	length	Most similar known cluster	MIBiG cluster type	MIBiG accession	% Similarity
NODE_36	86856	98.5089	Region 3.1	T1PKS, NRPS	1 - 86856	86856	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	50 %
NODE_97	58894	97.5162	Region 8.1	NRPS	1 - 39746	39746	aeruginoside 126B/aeruginoside 126A	NRP:Glycopeptide + Polyketide:Other + Saccharide:Hybrid/tailoring	BGC0000297	11 %
NODE_113	57256	97.8665	Region 10.1	NRPS,T1PKS,NRPS-like	1 - 55418	55418	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	62 %
NODE_198	48366	202.822	Region 18.1	lassopeptide	22813 - 45609	22797	NA	NA	-	-
NODE_198	48366	202.822	Region 20.1	T1PKS, NRPS-like,transAT-PKS-like	1 - 47734	47734	crocacin	NRP + Polyketide	BGC0000974	38 %
NODE_268	43207	96.1929	Region 32.1	T1PKS	11158 - 43207	32050	bartoloside2/bartoloside 3/ bartoloside 4	Other	BGC0001525	18 %
NODE_409	37235	99.9013	Region 47.1	T1PKS, NRPS	1 - 37235	37235	nostophycin	NRP + Polyketide	BGC0001029	27 %
NODE_415	36939	97.2136	Region 49.1	T1PKS, NRPS-like	1 - 36939	36939	nocuolin A	Other	BGC0001704	42 %
NODE_416	36936	99.7803	Region 50.1	NRPS,T1PKS,NRPS-like	1 - 36936	36936	malyngamide	NRP + Polyketide	BGC0001971	8 %
NODE_490	34421	96.7866	Region 60.1	hglE-KS,T1PKS	1 - 34421	34421	heterocyst glycolipids	Other	BGC0000869	100 %
NODE_585	31915	95.926	Region 68.1	NRPS	1 - 31915	31915	hassallidin C	NRP + Saccharide:Hybrid/tailoring	BGC0000369	18 %
NODE_680	29730	101.92	Region 77.1	terpene	3222 - 24187	20966	NA	NA	-	-
NODE_756	28198	97.6217	Region 86.1	lanthipeptide-class-v	1 - 28198	28198	NA	NA	-	-
NODE_851	26623	105.097	Region 90.1	NRPS	1 - 26623	26623	anabaenopeptin NZ857/ nostamide A	NRP	BGC0001479	100 %
NODE_958	24880	98.3543	Region 100.1	terpene	1 - 12954	12954	NA	NA	-	-
NODE_996	24317	98.5232	Region 103.1	redox-cofactor	3185 - 24317	21133	lankacidin C	NRP + Polyketide	BGC0001479	13 %
NODE_996	24317	98.5232	Region 112.1	NRPS	1 - 2364	2364	hassallidin C	NRP + Saccharide:Hybrid/tailoring	BGC0000369	25 %
NODE_1224	21952	121.756	Region 124.1	T1PKS, NRPS	1 - 21952	21952	puwainaphycin A/puwainaphycin B/puwainaphycin C/ puwainaphycin D	NRP + Polyketide	BGC0001125	30 %
NODE_1258	21587	100.506	Region 127.1	terpene	2004 - 21587	19584	geosmin	Terpene	BGC0000661	100 %
NODE_1294	21289	97.4493	Region 129.1	NRPS	1 - 21289	21289	NA	NA	-	-
NODE_1353	20800	124.546	Region 132.1	phosphonate	1 - 208	208	NA	NA	-	-
NODE_1364	20706	112.132	Region 135.1	NRPS-like	1 - 20706	20706	pelgipeptin	NRP	BGC0000403	37 %
NODE_1573	18812	305.754	Region 156.1	lanthipeptide-class-v, RRE-containing, spliceotide	1 - 18812	18812	NA	NA	-	-
NODE_1773	17224	102.796	Region 170.1	microviridin	1 - 17224	17224	microviridin	RiPP	BGC0000594	50 %
NODE_2000	15895	94.3449	Region 193.1	T1PKS	1 - 15895	15895	crochelin A	NRP + Polyketide	BGC0002001	7 %
NODE_2916	12106	98.3367	Region 250.1	NRPS-like,T1PKS	1 - 12106	12106	jamaicamide A/jamaicamide B/ jamaicamide C	NRP + Polyketide	BGC0001001	19 %
NODE_3702	9854	95.6489	Region 287.1	T1PKS	1 - 9854	9854	NA	NA	-	-
NODE_4136	8794	95.3816	Region 317.1	terpene	1 - 8794	8794	NA	NA	-	-
NODE_5351	6576	98.1219	Region 389.1	T1PKS	1 - 6576	6576	luminaolide	Polyketide	BGC0001656	13 %
NODE_6035	5680	100.528	Region 421.1	T1PKS	1 - 568	568	malyngamide C acetate	NRP + Polyketide	BGC0001970	20 %
NODE_6035	5680	100.528	Region 500.1	terpene	1 - 4004	4004	NA	NA	-	-
NODE_9883	3068	130.658	Region 545.1	T1PKS	1 - 3068	3068	1-heptadecene	Polyketide:Modular typer I	BGC0001164	100 %

Table 12. Putative biosynthetic gene clusters attributed using antiSMASH from Nostoc MAG-S93

Contig	contig length	contig coverage	Région	Cluster type (AntiSMASH)	Coordinates	length	Most similar known cluster	MIBiG cluster type	MIBiG accession	% Similarity
NODE_12	175094	53.717	Region 1.1	NRPS,T1PKS	67084 - 175094	108011	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	50 %
NODE_34	104508	56.1989	Region 4.1	terpene	66171 - 88528	22358	geosmin	terpene	BGC0000661	100 %
NODE_87	76146	52.2972	Region 9.1	hglE-KS,T1PKS	1 - 43245	43245	heterocyst glycolipids	Other	BGC0000869	100 %
NODE_90	75622	52.313	Region 11.1	NRPS-like,T1PKS,NRPS	20754 - 75622	54869	nostopeptolide A2	Polyketide + NRP:Cyclic depsipeptide	BGC0001028	62 %
NODE_120	70125	53.3808	Region 15.1	T1PKS,NRPS-like,transAT-PKS-like	16208 - 70125	53918	crocacin	NRP + Polyketide	BGC0000974	38 %
NODE_201	59862	69.3815	Region 22.1	NRPS	1 - 59862	59862	hassallidin C	NRP + Saccharide:Hybrid/tailoring	BGC0000369	56 %
NODE_315	50964	54.671	Region 31.1	microviridin	1 - 19246	19246	microviridin K	RiPP	BGC0000594	50 %
NODE_368	47520	55.0615	Region 38.1	phosphonate	1 - 39689	39689	NA	NA	-	-
NODE_382	46538	52.0665	Region 39.1	terpene	20032 - 40997	20966	NA	NA	-	-
NODE_500	41670	52.9015	Region 45.1	NRPS,T1PKS,NRPS-like	2461 - 4167	1707	malyngamide I	NRP + Polyketide	BGC0001971	8 %
NODE_560	39221	49.2008	Region 55.1	NRPS	1 - 33404	33404	aeruginoside 126B/aeruginoside 126A	NRP:Glycopeptide + Polyketide:Other + Saccharide:Hybrid/tailoring	BGC0000297	11 %
NODE_729	34234	54.125	Region 74.1	T1PKS	1 - 34234	34234	bartoloside 2/bartoloside 3/ bartoloside 4	Other	BGC0001525	18 %
NODE_881	31179	73.1629	Region 93.1	T1PKS	1 - 31179	31179	puwainaphycin A/puwainaphycin B/puwainaphycin C/ puwainaphycin D	NRP +Polyketide	BGC0001125	40 %
NODE_982	29081	53.3868	Region 105.1	redox-cofactor	1 - 21128	21128	lankacidin C	NRP + Polyketide	BGC0001100	13 %
NODE_1078	27496	55.5188	Region 113.1	NRPS	3393 - 27496	24104	jamaicamide A/jamaicamide B/ jamaicamide C	NRP + Polyketide	BGC0001001	7 %
NODE_1143	26516	54.7561	Region 119.1	NRPS	1 - 26516	26516	anabaenopeptin NZ857/ nostamide A	NRP	BGC0001479	100 %
NODE_1400	23019	52.1188	Region 136.1	terpene	239 - 23019	22781	NA	NA	-	-
NODE_1887	18263	48.7577	Region 166.1	T1PKS,NRPS	1 - 18263	18263	cryptophycin-327	NRP + Polyketide	BGC0000975	25 %
NODE_1924	18053	63.6398	Region 171.1	NRPS-like	1 - 18053	18053	pelgipeptin	NRP	BGC0000403	37 %
NODE_2147	16515	50.9377	Region 192.1	T1PKS	1 - 16515	16515	crochelin A	NRP + Polyketide	BGC0002001	7 %
NODE_2689	13359	54.9382	Region 227.1	NRPS	1 - 13359	13359	NA	NA	-	-
NODE_2763	12921	62.6087	Region 233.1	lanthipeptide-class-v	1 - 12921	12921	NA	NA	-	-
NODE_2978	11893	51.2665	Region 247.1	terpene	1 - 11893	11893	NA	NA	-	-
NODE_3273	10827	60.9827	Region 260.1	terpene	1 - 10827	10827	NA	NA	-	-
NODE_3542	9871	49.9163	Region 271.1	T1PKS	1 - 9871	9871	NA	NA	-	-
NODE_3563	9794	53.2863	Region 272.1	NRPS	1 - 9794	9794	cryptophycin-327	NRP + Polyketide	BGC0000975	25 %
NODE_4420	7452	49.982	Region 323.1	NRPS-like,T1PKS	1 - 7452	7452	1-heptadecene	Polyketide:Modular type I	BGC0001164	100 %
NODE_5683	5683	49.8612	Region 363.1	T1PKS	1 - 5683	5683	malyngamide C acetate	NRP + Polyketide	BGC0001970	20 %
NODE_8141	3211	50.9857	Region 430.1	NRPS-like	1 - 3211	3211	NA	NA	-	-

6. Discussion

The introduction of metagenomics in natural product research has the advantage to provide metabolic insights into microbial communities including novel species, therefore opening the gap to new insights into lichen symbiosis. However, the poor representability of the *Peltigera* species in the public databases posed particular challenges to draw up the protocol of the present study. As currently there is no available model to study *Nostoc*-mycobiont organisms, the main contribution of this work was therefore the development of *in silico* methodology and the selection of required bioinformatic tools to analyze the genetic and biosynthetic richness of the lichen-thalli dataset without the need of a reference model.

6.1. *De novo* assembly of MAGs

Whole genome sequencing of the entire lichen-thalli was performed to assess the genetic diversity of the *Peltigera* species, without the need for cultivation. Based on *de novo* assembly and binning, the protocol developed in this study succeeded in the generation of 261 draft genomes, including cyanobacteria, fungi as well as bacteria associated with lichen-specific phyla, with completeness higher than seventy percent. Considering the complexity of the *Peltigera* metagenomes, the implementation of a broad variety of bioinformatic tools was necessary to handle this dataset. Several steps of the protocol involved the implementation of more than one of them. Notably, the combination of *CONCOCT* and *metaBAT2* algorithms was required to proceed binning of the contigs. Overall, *metaBAT2* retained the highest number of medium-quality genomes, while *CONCOCT* performed better to detect and aggregate fungal contigs. Indeed, the binning process is still an imperfect science, especially for both prokaryotic and eukaryotic mixed communities. It is therefore common to use multiple tools for improving the accuracy of their results since, to date, no strategy outperformed the others. However, the comparison of their performances was a quite difficult task since the results produced by these tools could differ significantly. In addition, the huge amount of data generated had the potential to interfere with result interpretation. This observation highlighted the lack of available tools to deal with various datasets and multiple levels of complexity, particularly for performing binning assembly or quality assessment. Primarily, most of the currently available tools are still restricted to the analysis of procaryotes, thus urgently need to integrate some eukaryotes, especially fungi. Afterward, the refinement of sequences generated by *de novo* assembly was then a critical stage of the protocol since a large number of these sequences failed to resolve genomes with biological significance. This limitation came from the fact that *de novo* assembly and binning approaches were limited by the amount of information they can extract from short reads (Wu *et al.*, 2014; Sczyrba *et al.*, 2017). As result, even after the refinement of the binning dataset, most of the reconstructed genomes remained mostly fragmented and incomplete.

Despite the existing assembly tools' limitations, the method applied was able to resolve 53 high-quality MAGs. Since this study was intended for functional and metabolic investigations, the collection of MAGs, which was related to the *Nostoc* genus, provided the genetic material needed to screen bioactive metabolites and functional genes from lichen-associated cyanobionts of a range of *Peltigera* species.

6.2. *In silico* survey of MAGs for biosynthetic potential

The discovery of putative metabolic pathways was entitled to availability of high-quality assemblies (MAGs) and computational toolkits to identify metabolic gene clusters. Using the *antiSMASH* and *Palantir* algorithms, biosynthetic gene clusters were annotated, and the resulting secondary metabolites classes were predicted based on similarity to reference gene clusters and their known products in the MIBiG database.

Bioinformatic analysis of lichen cyanobiont BGCs offered the illustration of their potential to encode natural products of pharmaceutical significance through the detection of a large number of unknown and known BGCs. In silico, genome mining of Nostoc cyanobionts indicates the presence of a varied distribution of different, unique, and hybrid BGCs and changes in BGC contents among multiple strains. A single draft genome may contain 18 to 46 different BGCs, most of which are cryptic, while common types of BGCs that are found in all species were also uncovered. These ubiquitous BGCs were built from tree major producing genera of NPs, including non-ribosomal peptides (NRPs), polyketides (PKs) and their hybrids, and ribosomally synthesized and post-translationally modified peptides (RiPPs), which may play important antimicrobial roles in strain persistence, making them prospective therapeutic targets. The major enzymatic families detected, NRPS and T1-PKS, were especially abundant in the *Nostoc* genomes extracted from the metagenomes of *P.appalachensis*, *P.asiatica*, *P.borenquensis,* and *P.mikado*. Importantly, these *Nostoc* strains were precise whose could not be resolved at the species level. Since these *Nostoc* MAGs have been considered by *GTDB-tk* to be unique compared to the current NCBI database, they constituted a promising target for further comparative genomics. In this way, functional annotations supplied with the MAGs will constitute useful insights for deeper analysis of these cyanobionts. Furthermore, a great number of BGC predicted for the biosynthesis of a microviridin-like compound were identified. Microviridine possesses the greatest potential for new antibiotic and antitumor drugs. BGCs for RiPPs biosynthesis were revealed to be distributed widely in *Nostoc* from lichen *Peltigera* and will be a prolific producing resource for RiPP discovery.

In this framework, the detection of the BGCs was highly dependent on the composition of the draft genomes and the composition of the antiSMASH database because the genome mining tool has two major limitations. First, as cluster coordinates predicted by antiSMASH are based on a set of manually curated rules, the number of clusters detected depends on the size and the number of contigs. In the context of the metagenomic dataset, this means that when provided with fragmented draft genomes of low quality that splits the BGCs over multiple contigs, detection rules will certainly fail. In the case when BGCs cannot be fully characterized, the real number of the BGCs is overestimated, or their similarity to known clusters could be misattributed. Secondly, if the BGCs have not been described previously in the annotated database, rule-based genome mining tools will not be able to find them. Based on manual rules that may be too simplistic, the annotations of these multi-modular clusters are often incomplete. This consideration was encouraged in this context by the integration of the *Palantir* tool to ensure the optimisation of the annotations provided by *antiSMASH* for further analysis.

Finally, it should not be overlooked that, whereas cyanobionts of the *Nostoc* genus are increasingly being studied in recent years, a still tiny number of *Peltigera* mycobionts have been studied for the screening of new metabolites. However, some authors have already reported manifold cryptic BGCs to be new sources of natural products in different *Peltigera* species (Shukla *et al.*, 2010; Tanas *et al.*, 2010; Kampa *et al.*, 2013; Garg *et al.*, 2016. Difficulties in experimental and genetic manipulations of these species emphasize the need for establishing efficient metagenomic approaches applied to lichen-forming *Peltigera* species, which will further enable the survey of the biosynthetic potential of the six *Peltigera* species of this study.

Unlike, the present genome mining pipeline offered an efficient methodology to analyze BGCs of potential secondary metabolites from *Nostoc* symbiotic strains. If it is impossible today to reconstruct the complete and systematic assembly of the genomes arbored in these lichen communities from metagenomic data, *de novo* methods are already able to address important biological issues such as taxonomic characterization at the genus or species level and identification of functionally important genes. Certainly, the BGCs detected are better thought of as reflecting the potential rather than the real metabolic activities of the microorganisms. While *in silico* predictions offer a good starting point, we still needed to validate predictions by *in vitro* experiments. Once BGCs has been predicted to have a great biopotential, genomic information makes it possible to strategically engineer the heterologous expression of the desired compounds (Dhakal *et al*., 2021).

6.3. The future of Genome Mining

Accompanied by the current movement of massive acquisition of NGS datasets, the work of metagenomics is now facing a double challenge: first, the bioinformatic processing of these billions of sequences and heterogeneous information linked to them; second, relevant ecological or therapeutic biological interpretations of this data. Linking microbial biosynthetic genomic regions to their secondary metabolite products has largely been done manually on small numbers of strains (e.g., Kaweewan *et al.*, 2017; Xu *et al.,* 2018), but traditional screening approaches for the discovery of new natural products are no longer profitable. Current genomic studies have diversified towards the emergence of new 'omic' technologies (metagenomics, transcriptomics, proteomics, metabolomics, *etc*.) that break down biological information from the different levels of organization and functioning of a biological system or community. The challenge today is to integrate these different disciplines, in addition to experimental data such as synthetic approaches and sets of reference library MS/MS spectra, to correlate the data. Multiple efforts are therefore being made to standardize data, including meta-data (e.g. sequencing coverage, paired-end incongruence, *etc*.) and editorial vocabulary (policy) while providing guidance to ensure that these data are all comparable in terms of quality. Perspectives have also appeared thanks to the use of increasingly innovative statistical tools (sensitivity and detection capacity improvements in sequencing, annotation, assembly, etc.), in conjunction with the development of additional cutting-edge technologies such as machine learning and smart data analytics. These advances have permitted the prediction of metabolites still unknown in the databases and boosted multi-omics design. Over time, the improvements have granted genome mining studies the potential to reduce rediscovery rates of known metabolites, guide experimental work towards the most promising candidates with high impact in industries, and identify enzymatic pathways that enable their biosynthetic production (Van der Hooft *et al.*, 2020).

7. Perspectives

In a bioinformatics context, one of the important points of this thesis was also the exploration of technological possibilities to meet the objectives of the study. According to the exponential development of bioinformatics algorithms, a series of tools was reviewed during this work, and some examples were considered in the present perspective.

The main restriction of metagenomic using second-generation sequencing is the limited length of reads, which inevitably reduced the resolution possibilities of the assembler (Saary *et al.,* 2020). Long reads data can be used to 'bridge' short reads from the same genomic region and by consequent, to overcome the limitations of genome assembly of short reads. Nowadays, as third-generation sequencing platforms (see Table 1), continue to improve in accuracy (Sereika *et al.*, 2022), hybrid assemblies combining the number of short reads yielded by second-generation technologies, and the length of reads provided by third-generation technologies will become the norm. Applied to this study, the hybrid assembly will significantly improve the resolution of the MAGs, in terms example of genes completeness, closely related strains differentiation, and common repeat elements such as 16S rRNA operons. The loss of ribosomal S16 RNA operon during the binning process as previously described by authors Cornet *et al.* (2018) and Nelson *et al.* (2020), is in fact a significant drawback since these elements are commonly used for taxonomic identification. Additionally, supplementary markers *rbcLX* and *trnL* will constitute in the MAGs of *Nostoc* sufficient genomic data for their identification at the species level (Gagunashvili and Andrésson, 2018). Naturally, by incorporating extrinsic information, such as a database of genes or reference genomes, it will be possible to achieve a completely different descriptive resolution, and to further characterize the functional potential of these systems.

Whereas assemblies are constructed from graphs, new binning tools take advantage of the graph information prior to consensus linear contigs. Using such tools like *GraphBin* (Mallawaarachchi *et al.*, 2020), could potentially maximize the number of reads retrieved in MAGs. The implementation of such binners remains to be evaluated on the *Peltigera* communities. Afterwards, curation of the bins set based on visual comparison of the bins as proposed by *Anvio* software (Murat Eren *et al.,* 2015) is a more accurate method to manually refine the bins compared to consideration of estimated completeness and contamination only. As a major point of this end, high quality MAGs of the lichen-forming mycobiont may be retained and amenable to functional analysis them to. Then, contigs of potential rarer strains that cannot be binned into MAGs are still a catalogue of diversity that can be mined in the same way as MAGs.

In addition, although more and more tools offer unsupervised methods, the machine learning algorithms they return to must further enrich their performance through training on sets of diverse populations. Thus to improve the automatic binning of MetaBAT2 to provide better results on lichens, it would be wise to train this machine learning algorithm on data sets containing both bacteria and eukaryotes.

Among the most promising, several orphan BGCs were assigned to RiPPs gene cluster types. Considering their great interest in medicines, the future research should focus on deep genome sequencing and BGC predication methods to precisely determine the presence and gene content characteristics of these specific BGCs in a genome. Once these BGCs well characterized, genomic information makes it possible to strategically engineer their heterologous expression.

Finally, as the reproducibility and transposability are a key challenge for the future of metagenomics, and already an point of contention, it will be important in the next step to communicate a script like NextFlow (https://www.nextflow.io) allowing the implementation of the entire bioinformatics pipeline presented. This would help to overcome the difficulties inherent in the perpetual evolution of operating systems as well as the future updates of versions specific to the software used.

Conclusion

Lichens are organisms with a major ecological role demonstrated by a unique chemical profile. The ability of lichens to synthesize a wide range of bioactive secondary metabolites gives rise to many biotechnological interests for which various pharmaceutical activities are particularly promising. However, only a limited number of secondary metabolites from lichens have been used for biological applications so far. The main reasons are their slow growth in nature and experimental limitations that make it difficult to establish lichen cultures. As a result, a very limited number of genomes of lichen-associated organisms are currently available in public databases and the mechanisms underlying their molecular biosynthesis are still poorly understood.

This study presents the *de novo* metagenomic survey of a large collection of cyanolichens associated with six *Peltigera* species. About the current need for new drug leads, this study focuses on the biosynthetic potential of secondary metabolites of lichen-forming *Peltigera* symbionts. In this paper, a framework was designed to recover the diversity of secondary metabolite biosynthesis at the symbiont level from previously assembled *de novo* metagenomic reads.

These bioinformatics developments can be grouped into two steps: *(i) de novo* metagenomic assembly and sequence profiling, which divides the metagenome sequences into bins of the same taxonomic origin; *(ii) in silico* genome mining for biosynthetic gene clusters, which studies the genome in terms of genes related to biosynthetic pathways, to assess the potential variety of natural products synthesized.

The first objective was high-quality genome reconstruction of the metagenome-assembled genomes (MAGs) of the cyanolichen community. The complexity of lichen metagenomes makes *de novo* assembly of symbiotic partners algorithmically challenging, as there are large differences in abundance between individual taxa in the prokaryote and eukaryote phylogroups. Despite this, the *de novo* assembly pipeline allowed the reconstruction of high-quality draft genomes. Of the 55 symbiotic communities sequenced, 54 MAGs of the cyanobionts, genus *Nostoc*, were obtained. This was made possible by the integration and strengths of multiple bioinformatics tools and iterative refinement methods. By combining the availability of high-quality MAGs, this study provided, as a next step, a first insight into the formation of secondary metabolites in these collections of Nostoc associated to *Peltigera* species. *In silico* genome mining using advanced bioinformatics tools revealed 1054 BGCs, most of which are 'orphans' and cannot be expressed under standard laboratory culture conditions, suggesting a potential untapped production of many unknown metabolites. In addition, major natural product pathways, including NRPS/PKS groups and RiPPs, were found in all *Nostoc* strains. BGCs of these major pathways are known for the biosynthesis of a variety of natural products, including antibacterial and antitumor, and were particularly detected in the genomes of *Nostoc* associated to *P.borenquensis*. This study showed that *Peltigera* cyanobionts were capable of producing a wide range of bioactive secondary metabolites with potential therapeutic and biotechnological applications.

By providing high-quality MAGs, this framework opens the way for downstream studies based on metabolic and functional approaches, but also any other genome exploration interest. Expanding the understanding of biosynthetic genes in lichen communities can help to better understand the underlying biosynthetic mechanisms to consider their applicability to the pharmaceutical industry. Although the metagenomic framework has improved the knowledge of *Peltigera* cyanobionts, this study has also highlighted some limitations of the framework, such as the availability of reference genomes, the difficulty of dealing with highly complex metagenomic data (in both prokaryotes and eukaryotes) and, finally, the error-prone data associated with *de novo* metagenomic approaches and *in silico* genomic exploration.

Bibliography

Albarano L, Esposito R, Ruocco N, Costantini M. Genome mining as new challenge in natural products discovery. *Mar Drugs*. 2020;18(4). doi:10.3390/md1804019

A R. Cyanolichens: Nitrogen metabolism. *Rai A, Bergman B, Rasm U Cyanobacteria symbiosis*. 2002:97-115.

Aschenbrenner IA, Cernava T, Berg G, Grube M. Understanding microbial multi-species symbioses. *Front Microbiol*. 2016;7(FEB). doi:10.3389/fmicb.2016.00180

Asplund J, Wardle DA. How lichens impact on terrestrial community and ecosystem properties. *Biol Rev*. 2017;92(3):1720-1738. doi:10.1111/brv.12305

Aziz RK, Bartels D, Best A, et al. The RAST Server: Rapid annotations using subsystems

Bertrand RL, Sorensen JL. Lost in Translation: Challenges with Heterologous Expression of Lichen Polyketide Synthases. *ChemistrySelect*. 2019;4(21):6473-6483. doi:10.1002/slct.201901762

Bernardo P, Albina E, Eloit M, Roumagnac P. Métagénomique virale et pathologie: Une histoire récente. *Medecine/Sciences*. 2013;29(5):501-508. doi:10.1051/medsci/2013295013

Boustie J, Grube M. Lichens—a promising source of bioactive secondary metabolites. *Plant Genet Resour Charact Util*. 2005;3(02):273-287. doi:10.1079/PGR200572

Burgos-Toro A, Dippe M, Vásquez AF, Pierschel E, Wessjohann LA, Fernández-Niño M. Multi-Omics Data Mining: A Novel Tool for BioBrick Design. *Synth Genomics - From BioBricks to Synth Genomes*. 2022. doi:10.5772/intechopen.101351

Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: Increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res*. 2021;49(D1):D639-D643. doi:10.1093/nar/gkaa978

Blin K, Shaw S, Kloosterman AM, et al. AntiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Res*. 2021;49(W1):W29-W35. doi:10.1093/nar/gkab335

Calchera A, Grande FD, Bode HB, Schmitt I. Biosynthetic Gene Content of the 'Perfume Lichens' Evernia prunastri and Pseudevernia furfuracea. *Molecules*. 2019;24(1). doi:10.3390/molecules24010203

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2020;36(6):1925-1927. doi:10.1093/bioinformatics/btz848

Cernava T, Erlacher A, Aschenbrenner IA, et al. Deciphering functional diversification within the lichen microbiota by meta-omics. *Microbiome*. 2017;5(1):82. doi:10.1186/s40168-017-0303-5

do Amaral SC, Monteiro PR, da Silva Pinto Neto J, et al. Current Knowledge on Microviridin from Cyanobacteria. *Mar Drugs*. 2021;19(1). doi:10.3390/md19010017

Eldjarn GH, Ramsay A, Hooft JJJ Van Der, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput Biol*. 2021;17(5 May). doi:10.1371/journal.pcbi.1008920

F. FJ. The lichen as an ecosystem: observation and experiment. *Lichenol Prog Probl*. 1976:385-406.

Gagunashvili AN, Andrésson ÓS. Distinctive characters of Nostoc genomes in cyanolichens. *BMC Genomics*. 2018;19(1). doi:10.1186/s12864-018-4743-5

Giglio S, Jiang J, Saint CP, Cane D, Monis PT. Isolation and Characterization of the Gene Associated with Geosmin Production in Cyanobacteria. http://pubs.acs.org.

Grimm M, Grube M, Schiefelbein U, Zühlke D, Bernhardt J, Riedel K. The Lichens' Microbiota, Still a Mystery? *Front Microbiol*. 2021;12(March):1-25. doi:10.3389/fmicb.2021.623839

Grube M, Berg G, Andrésson ÓS, Vilhelmsson O, Dyer PS, Miao VPW. 9 Lichen Genomics: Prospects and Progress. 2014.

Grujić V, Thesis MS. UNIVERSITY OF LJUBLJANA BIOTECHNICAL FACULTY STUDY OF BIOTECHNOLOGY APPLICATION OF antiSMASH TOOL FOR IDENTIFICATION AND ANALYSIS OF TYPE II POLYKETIDE GENE CLUSTERS.

Goga M, Elečko J, Marcinčinová M, Ručová D, Bačkorová M, Bačkor M. Lichen Metabolites: An Overview of Some Secondary Metabolites and Their Biological Potential. 2018:1-36. doi:10.1007/978-3-319-76887-8_57-1

Gonnet D, Gonnet O, Van Haluwyn C, Méral J-P. *A La Découverte Des Lichens: Stage d'initiation*. Paris; 2017.

Hager A, Brunauer G, Türk R, Stocker-Wörgötter E. Production and bioactivity of common lichen metabolites as exemplified by Heterodea muelleri (Hampe) Nyl. *J Chem Ecol*. 2008;34(2):113-120. doi:10.1007/s10886-007-9408-9

Hofmann FMP, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy AC. AMBER: Assessment of Metagenome BinnERs. Gigascience. 2018;7(6). doi:10.1093/gigascience/giy069

Honegger R. Simon Schwendener (1829-1919) and the Dual Hypothesis of Lichens. Bryologist. 2000;103(2):307-313.

Hooft JJJ Van Der, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev. 2020;49(11):3297-3314. doi:10.1039/d0cs00162g

Iskandar IK, Syers JK. Solubility of Lichen Compounds in Water: Pedogenetic Implications. Lichenol. 1971;5(1-2):45-50. doi:10.1017/S0024282971000082

Kaasalainen U, Fewer DP, Jokela J, Wahlsten M, Sivonen K, Rikkinen J. Cyanobacteria produce a high variety of hepatotoxic peptides in lichen symbiosis. Proc Natl Acad Sci U S A. 2012;109(15). doi:10.1073/pnas.1200279109

Kampa A, Gagunashvili AN, Gulder TAM, et al. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. Proc Natl Acad Sci U S A. 2013;110(33). doi:10.1073/pnas.1305867110

Kautsar SA, Blin K, Shaw S, et al. MIBiG 2.0: A repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2020;48(D1):D454-D458. doi:10.1093/nar/gkz882

Kautsar SA, Duran HGS, Blin K, Osbourn A, Medema MH. PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 2017;45(W1):W55-W63. doi:10.1093/nar/gkx305

Kealey JT, Craig JP, Barr PJ. Identification of a lichen depside polyketide synthase gene by heterologous expression in Saccharomyces cerevisiae. Metab Eng Commun. 2021;13. doi:10.1016/j.mec.2021.e00172

Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. *Nat Rev Microbiol*. 2019;17(3):167-180. doi:10.1038/s41579-018-0121-1

.Kim W, Liu R, Woo S, et al. Linking a gene cluster to atranorin, a major cortical substance of lichens, through genetic dereplication and heterologous expression. MBio. 2021;12(3). doi:10.1128/mBio.01111-21

Lapidus AL, Korobeynikov AI. Metagenomic Data Assembly – The Way of Decoding Unknown

Microorganisms. Front Microbiol. 2021;12. doi:10.3389/fmicb.2021.613791

Leão T, Wang M, Moss N, et al. A Multi-Omics Characterization of the Natural Product Potential of Tropical Filamentous Marine Cyanobacteria. Mar Drugs. 2021;19(1). doi:10.3390/md19010020

LePogam P. Analyses de lichens par spectrométrie de masse : déréplication et histolocalisation. 2016.

Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6. doi:10.1038/srep19233

Lücking R. Stop the Abuse of Time! Strict Temporal Banding is not the Future of Rank-Based Classifications in Fungi (Including Lichens) and Other Organisms. CRC Crit Rev Plant Sci. 2019;38(3):199-253. doi:10.1080/07352689.2019.1650517

Lutzoni F, Miadlikowska J. Quick guide: Lichens. Curr Biol. 2009;19(13):R502-R503. doi:10.1016/j.cub.2009.04.034

Magain N, Truong C, Goward T, et al. Species delimitation at a global scale reveals high species richness with complex biogeography and patterns of symbiont association in peltigera section peltigera (Lichenized ascomycota: Lecanoromycetes). Taxon. 2018;67(5):836-870. doi:10.12705/675.3

Miadlikowska J, Lutzoni F. PHYLOGENETIC REVISION OF THE GENUS PELTIGERA (LICHEN-FORMING ASCOMYCOTA) BASED ON MORPHOLOGICAL, CHEMICAL, AND LARGE SUBUNIT NUCLEAR RIBOSOMAL DNA DATA. Int J Plant Sci. 2000;161(6):925-958.

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. Bioinformatics. 2016;32(7):1088-1090. doi:10.1093/bioinformatics/btv697

Magain N, Miadlikowska J, Goffinet B, Serusiaux E, Lutzoni F. Macroevolution of specificity in cyanolichens of the genus Peltigera section Polydactylon (Lecanoromycetes, Ascomycota). In: Systematic Biology. Vol 66. Oxford University Press; 2017:74-99. doi:10.1093/sysbio/syw065

Magain N, Miadlikowska J, Mueller O, et al. Conserved genomic collinearity as a source of broadly applicable, fast evolving, markers to resolve species complexes: A case study using the lichen-forming genus Peltigera section Polydactylon. Mol Phylogenet Evol. 2017;117:10-29. doi:10.1016/j.ympev.2017.08.013

Mallawaarachchi VG, Wickramarachchi AS, Lin Y. Improving metagenomic binning results with overlapped bins using assembly graphs. Algorithms Mol Biol. 2021;16(1). doi:10.1186/s13015-021-00185-6

Mallawaarachchi V, Wickramarachchi A, Lin Y. GraphBin: Refined binning of metagenomic

contigs using assembly graphs. Bioinformatics. 2020;36(11):3307-3313. doi:10.1093/bioinformatics/btaa180

Medema MH, Kottmann R, Yilmaz P, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11(9):625-631. doi:10.1038/nchembio.1890

Meiser A, Otte J, Schmitt I, Grande FD. Sequencing genomes from mixed DNA samples - Evaluating the metagenome skimming approach in lichenized fungi. Sci Rep. 2017;7(1). doi:10.1038/s41598-017-14576-6

Meunier L, Tocquin P, Cornet L, et al. Palantir: A springboard for the analysis of secondary metabolite gene clusters in large-scale genome mining projects. Bioinformatics. 2020;36(15):4345-4347. doi:10.1093/bioinformatics/btaa517

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. Bioinformatics. 2016;32(7):1088-1090. doi:10.1093/bioinformatics/btv697

Miadlikowska J, Lutzoni F. Phylogenetic classification of peltigeralean fungi (Peltigerales, Ascomycota) based on ribosomal RNA small and large subunits. Am J Bot. 2004;91(3):449-464. doi:10.3732/ajb.91.3.449

Miadlikowska J, Lutzoni F. PHYLOGENETIC REVISION OF THE GENUS PELTIGERA (LICHEN-FORMING ASCOMYCOTA) BASED ON MORPHOLOGICAL, CHEMICAL, AND LARGE SUBUNIT NUCLEAR RIBOSOMAL DNA DATA. Int J Plant Sci. 2000;161(6):925-958.

Nash TH. Lichen Biology. 2nd ed. (Press CU, ed.).; 2008.

Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod. 2016;79(3):629-661. doi:10.1021/acs.jnatprod.5b01055

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;27(5):824-834. doi:10.1101/gr.213959.116

Ochman H, Worobey M, Kuo CH, et al. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biol. 2010;8(11). doi:10.1371/journal.pbio.1000546

Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12. doi:10.1186/1471-2105-12-385

Palazzotto E, Weber T. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. Curr Opin Microbiol. 2018;45:109-116. doi:10.1016/j.mib.2018.03.004

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the

quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043-1055. doi:10.1101/gr.186072.114

Ponsero AJ, Hurwitz BL, Magain N, Miadlikowska J, Lutzoni F, U'Ren JM. Cyanolichen microbiome contains novel viruses that encode genes to promote microbial metabolism. ISME Commun. 2021;1(1). doi:10.1038/s43705-021-00060-w

Ranković B, Ranković D, Kosanić M, Marić D. Antioxidant and antimicrobial properties of the lichens Anaptychya ciliaris, Nephroma parile, Ochrolechia tartarea and Parmelia centrifuga. Cent Eur J Biol. 2010;5(5):649-655. doi:10.2478/s11535-010-0043-z

Resl P, Bujold AR, Tagirdzhanova G, et al. Large differences in carbohydrate degradation and transport potential among lichen fungal symbionts. Nat Commun. 2022;13(1). doi:10.1038/s41467-022-30218-6

Rikkinen J. Algal and Cyanobacteria Symbioses Symbiotic Cyanobacteria in Lichens. Algal and Cyanobacteria Symbioses. 2017;Chapter 5:147-167. www.worldscientific.com.

Rikkinen J. Cyanolichens. Biodivers Conserv. 2015;24(4):973-993. doi:10.1007/s10531-015-0906-8

Rikkinen J. Molecular studies on cyanobacterial diversity in lichen symbioses. MycoKeys. 2013;6:3-32. doi:10.3897/mycokeys.6.3869

Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nat Rev Microbiol. 2018;16(12):731-744. doi:10.1038/s41579-018-0075-3

Roullier C. Recherche de mycosporines et de dérivés aminés lichéniques d'intérêt pour les cancers photoinduits . Etude phytochimique d'un lichen marin : Lichina pygmaea (Lightf.) C. 2010:258.

Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. Front Genet. 2017;8(MAR). doi:10.3389/fgene.2017.00023

Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. Genome Biol. 2020;21(1). doi:10.1186/s13059-020-02155-4

Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4. doi:10.1186/s40168-016-0154-5

Scherlach K, Hertweck C. Mining and unearthing hidden biosynthetic potential. Nat Commun. 2021;12(1). doi:10.1038/s41467-021-24133-5

Sczyrba A, Hofmann P, Belmann P, et al. Critical Assessment of Metagenome Interpretation -

A benchmark of metagenomics software. Nat Methods. 2017;14(11):1063-1071. doi:10.1038/nmeth.4458

Shaffer M, Borton MA, McGivern BB, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48(16):8883-8900. doi:10.1093/nar/gkaa621

Shih PM, Wu D, Latifi A, et al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A. 2013;110(3). doi:10.1073/pnas.1217107110

Shishido TK, Wahlsten M, Laine P, Rikkinen J, Lundell T, Auvinen P. Microbial Communities of Cladonia Lichens and Their Biosynthetic Gene Clusters Potentially Encoding Natural Products.

Shvarev D, Nishi CN, Maldener I. Glycolipid composition of the heterocyst envelope of Anabaena sp. PCC 7120 is crucial for diazotrophic growth and relies on the UDP-galactose 4-epimerase HgdA. Microbiologyopen. 2019;8(8). doi:10.1002/mbo3.811

Sierra MA, Danko DC, Sandoval TA, et al. The Microbiomes of Seven Lichen Genera Reveal Host Specificity, a Reduced Core Community and Potential as Source of Antimicrobials. Front Microbiol. 2020;11. doi:10.3389/fmicb.2020.00398

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351

Simon A, Goffinet B, Magain N, Sérusiaux E. High diversity, high insular endemism and recent origin in the lichen genus Sticta (lichenized Ascomycota, Peltigerales) in Madagascar and the Mascarenes. Mol Phylogenet Evol. 2018;122:15-28. doi:10.1016/j.ympev.2018.01.012

Singh G, Calchera A, Schulz M, et al. Climate-specific biosynthetic gene clusters in populations of a lichen-forming fungus. Environ Microbiol. 2021;23(8). doi:10.1111/1462-2920.15605

Society E. Removal of Lichen Secondary Metabolites Affects Food Choice and Survival of Lichenivorous Moth Larvae Author (s): Heikki Pöykkö , Marko Hyvärinen and Martin Bačkor Published by : Wiley on behalf of the Ecological Society of America Stable URL : http://w. 2017;86(10):2623-2632.

Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1). doi:10.1186/s13321-020-00424-9

Soueidan H, Nikolski M. Machine learning for metagenomics: methods and tools. 2015. http://arxiv.org/abs/1510.06621.

Stocker-Wörgötter E. Metabolic diversity of lichen-forming ascomycetous fungi: Culturing, polyketide and shikimate metabolite production, and PKS genes. Nat Prod Rep. 2008;25(1):188-200. doi:10.1039/b606983p

T. Resl PSDE& TGS. Evolutionary biology of lichen symbioses. New Phytol. 2022;234:1152-1566.

Van Haluwyn C, Asta J, Gaveriaux J-P. Guide Des Lichens de France: Lichens Des Arbres. Gavériaux. (Belin, ed.).; 2013.

Vollmers J, Wiegand S, Kaster AK. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! PLoS One. 2017;12(1). doi:10.1371/journal.pone.0169662

Walz SEW. N E U R O M E T H O D S. http://www.springer.com/series/7657.

Wang Y, Geng C, Yuan X, Hua M, Tian F, Li C. Identification of a putative polyketide synthase gene involved in usnic acid biosynthesis in the lichen Nephromopsis pallescens. PLoS One. 2018;13(7). doi:10.1371/journal.pone.0199110

W. LD. The Absorption and Release of Water by Lichens. Vol 25. (Lichenologica B, ed.). Bibliotheca Lichenologica; 1987.

Yang C, Chowdhury D, Zhang Z, et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Comput Struct Biotechnol J. 2021;19:6301-6314. doi:10.1016/j.csbj.2021.11.028

Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes-a review. Nat Prod Rep. 2016;33(8):988-1005. doi:10.1039/c6np00025h

Appendices

Annex 1. Vocabulary used in lichenology. [Copy from Gonnet et al., 2017].

- **Apothecia**: the most common sexual reproduction structure of the lichenic fungal partner, related to the production of spores. These ascoma are differentiated by the mycosymbiote of sectional discolichens, sessile or stipitate, containing the hymenium exposed to the open air.
- **Fibrilla**: Very thin filament, +/- lying on the surface that supports it. In the Usnea, the fibrils (when they exist) are the short filamentous outgrowths, rarely more than 1 cm long and 0.2 mm Ø, with a central axis not connected to that of the twig which carries them, not or barely divided, inserted perpendicular around the secondary and terminal branches. Note: Very short fibrils are called spinules.
- **Cephalodia**: Small well-delineated formation containing a secondary photosymbiote different from the primary photosymbiote that dominates the thallus. Often pockets of cyanobacteria that form on top of a chlorolichen, normally black in color. Often pockets of cyanobacteria that form on top of a chlorolichen, normally black in color. Headaches are present in a few genera: Lobaria, Nephroma, Peltigera, Solorina, Stereocaulon, etc.
- **Cilia**: Filiform elements, often darker, consisting of the extensions of several hyphae adjoined (not to be confused with hair). Filiform elements, often darker, consisting of the extensions of several hyphae adjoined (not to be confused with hair).
- **Chlorolichens**: a lichen that has a green alga for its photobiont.
- **Cortex**: The protective outer wall of the thallus, composed entirely of fungal tissue. Lichens may have two cortices (upper and lower), a single cortex or no cortex at all, depending on growth form. Below the cortex is the photobiont.
- **Crustose**: A lichen growth form distinguished by the thallus being tightly adhered to the substrate at all points. Crustose lichens do not have a lower cortex, exposing the hyphae to the substrate. It is impossible to remove a crustose lichen from its substrate without impacting the substrate in some way.
- **Cyanolichen**: a lichen that has cyanobacteria for its primary photobiont.
- C**yphelle**: Small cortical depression, cut-like, on the underside of the thallus of lichens of the genus Sticta where they are +/- masked by the tomentum.
- **Foliose**: A lichen growth form distinguished by a relatively flat, leaf-like thallus. Foliose lichens have an upper and lower cortex, making it easy to identify an upper and lower thallus surface.
- F**ruticose**: A lichen growth form distinguished by a tufted, hanging or stalked thallus. Fruticose lichens have a single, continuous cortex that wraps around the thallus branches, making it difficult to discern an upper and lower surface.
- **Hapter** : is a fixation spike. It is an organ used to attach the base of a thallus to the substrate. Unlike a root, the hapter does not perform any function of absorption, either of water or nutrient.
- **Hyphae** : Microscopic filamentous strands constituting the fungal cells that make up the thallus of a lichen and encapsulate the algal cells. Hyphae have a rigid wall consisting of glucans, glycoproteins, proteins and chitin, they have a moving flow of cytoplasm allowing apical growth; they represent 80 to 90% of the biomass of a lichen.
- **Gonidia**: In lichens, name formerly assigned to chlorophyll cells living in association with hyphae. These gonidias, before the designs of Schwendener (1867 and 1869), were considered as elements involved in the reproduction of the lichen. These gonidia are actually photosymbiotes: green algae (phycosymbiotes) or cyanobacteria (cyanosymbiotes).
- **Isidia**: Small cortical growths, normally found on the top-side or outer cortex of the lichen, developed by the thallus. They contain cells of the main photosymbiote, cells of the mycosymbiote and are surrounded by a tight layer of hyphae, prolongation of the upper cortex, and consequently generally concolored to the thallus.

- **Lichenized**: a fungus, alga, or cyanobacterium that is in a lichen partnership.
- **Lobe**: a flattened branch, generally found on foliose lichens.
- **Lobule**: a vegetative means of propagation for lichens; small lobe-like two-sided propagules that break off and reestablish elsewhere.
- **Macrolichen**: a foliose or fruticose lichen; physical features are seen with the naked eye.
- **Medulla**: a generally loose layer of hyphae below the cortex and photosynthetic layer, often containing air spaces and crystals of lichen substances.
- **Microlichen**: a crustose or squamulose lichen; physical features are difficult to see with a naked eye and a microscope is required for identification.
- **Propagule**: a reproductive or dispersal body containing both the algal and fungal components, i.e. the photobiont and mycobiont.
- **Perithecium** (perithecia): One type of fruiting structure produced by the fungal component of the lichen. A perithecium is flask-shaped (compare with apothecium) and often embedded the thallus, making it somewhat inconspicuous. A small hole at the top of the perithecium releases spores, which allow for sexual reproduction.
- **Rhizines**: Linear or narrow root-like appendages that protrude from the lower thallus surface (compare with cilia) and attach to the substrate.
- **Soredia** (-**Soralia**): a vegetative means of propagation for lichens. Both fungus and alga are intertwined into a granule-like mass that occurs on top of the cortex and on the margins. Some lichens have structures called soralia that produce soredia.
- **Spore**: in lichens, produced by the mycobiont only for sexual reproduction; a tiny uni- or multi-cellular structure that gives rise to another fungus; no photobiont is included within the spore.
- **Squamulose**: A lichen growth form distinguished by small, overlapping thallusunits or scales. Squamulose lichens are not as tightly appressed to the substrate ascrustose lichens but are more appressed than foliose lichens. These lichens have anupper cortex but may or may not have a lower cortex.
- **Squamules** : small foliose lichen structures that are attached to their substrate by one end, like a shingle; several of these structures will comprise a lichen.
- **Thallus** (thalli) : the vegetative body of the lichen, composed of both fungus and alga
- **Tomentum** (tomentose): colorless hyphae that look like short fuzz or hairs on the outside of the lichen.
- **Voucher specimen**: one collected as evidence of the occurrence of a particular species in a place, or to support another type of data such as DNA or TLC or anatomical slides.

Annex 2. Phylogeny of the lichen-forming genus Peltigera (mycobiont, i.e., fungal partner). [Copy from Magain *et al.*, 2017]

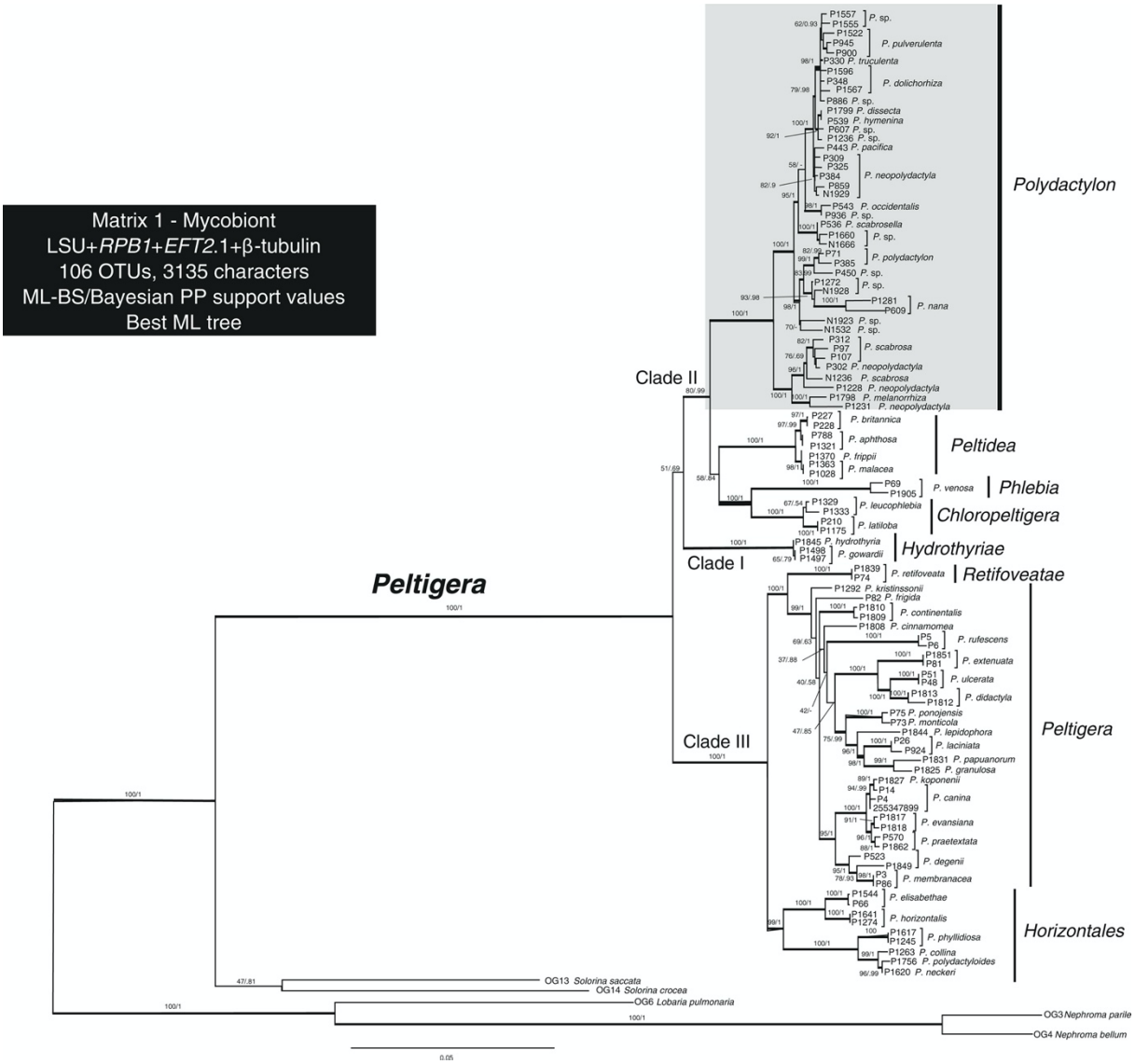


Fig: Phylogeny of the lichen-forming genus Peltigera (mycobiont, i.e., fungal partner). Most likely tree derived from an ML search using Matrix 1 (Table 1, Fig. 1), consisting of 106 OTUs representing 64 species from all (eight) sections of Peltigera and five outgroup species selected from the genera Solorina (Peltigeraceae), Lobaria (Lobariaceae), and Nephroma (Nephromataceae). The tree was rooted according to Miadlikowska and Lutzoni (2004). Values associated with each internode represent ML bootstrap support (ML-BS; before slash) and Bayesian posterior probabilities (PP; after slash). Thick internal branches represent internodes with ML-BS ⩾ 70 and PP ⩾ 0.95. Vertical bars delimit sections of the genus Peltigera as circumscribed by Miadlikowska and Lutzoni (2000). A gray box highlights the focal section–-Polydactylon.

Annex 3. Genome Binning: additional knowledge

The recovery of strain-specific genotypes from environmental communities is a complex task and the major challenge in metagenomic research. Consequently, a variety of binning algorithms have been developed to attempt the clustering of the pre-assembled fragments, contigs or scaffolds, from the same, or closely related, species or higher taxonomic levels (Damayanthi *et al.*, 2017; Yi *et al*., 2020). The binned fragments are then used to construct the draft, or sometimes complete, genomes of the different strains found in the metagenomic sample. Depending on the identification of these candidate MAGs, subsequent bioprospecting such as taxonomic assignment and functional analysis can then be done in the bins instead of individual contigs (Breitweiser *et al*., 2019).

Supplementary Table lists the main methodological features of varied metagenome binning approaches. Both supervised and unsupervised machine learning methods are employed in the binning process. Since methods relying on available current datasets will miss important properties of samples, most existing computational tools for binning utilize hybrid strategies from unsupervised methods. Such an approach group sequences into unlabeled bins without consideration of external reference or taxonomic information by considering various measures of similarity such as guanine-cytosine content, oligomer frequencies, read depth or contig coverage. These latter measures differ from one organization to another and therefore allow automatic sorting to be carried out. Despite the development of many tools for binning, there is no single best choice, and the accuracy of results is still difficult to gauge (Nelson *et al.*, 2020). Also of note, the current literature only collects very few pipelines allowing the binning of metagenomes composed of both a predominant eukaryote and bacteria, as is the case in this present study of lichenic samples. Consequently, combining several current binning software to complement each other is a common strategy, and has gradually become the mainstream option.

Supplementary Table. Key methodological features of the common genome binning approaches.

Method	Starting point	Clustering methods	Negatives	Positives	Computational Resources
Nucleotide composition (NC)	Oligonucleotide frequency matrix and %G+C-based screening.	HCL, correlation-based network graph and emergent self-organization maps (ESOM).	(i) More efficent for the genomes with skewed nucleotide composition patterns. (ii) Less efficient in differentiating between closely related genotypes. (iii) Depends on the visualization and manual inspection of bins and therefore are not suitable for very large assemblies representing complex environments.	(i) Individual metagenome assemblies or samples where populations do not change over time can be used.	(i) R packages: qgraph, i graph, pv-clust (Suzuki & Shimodaira, 2006) (ii) tetramerFreqs (Dick *et al.*, 2009) (iii) Databionic ESOM tools (Ultsch & Mörchen, 2005) (iv) 2T-binning (Saeed *et al.*, 2011)
Nucleotide composition and abundance (NCA)	A composite distance matrix from oligonucleotide frequency matrix and coverage.	K-medioids clustering, Gaussian mixture models, and expectation and maximization algorithm.	(ii), (iv) Require multiple samples for better performance, and therefore are associated with cost, time, and computational resources.	(i), (ii) Improved contig binning than NC method	(i) MetaBAT (Kang *et al.*, 2015) (ii) CONCOCT (Alneberg *et al.*, 2014) (iii) MaxBin (Wu *et al.*, 2014) (iv) GroopM (Imelfort *et al.*, 2014) (v) Databionic ESOM tools (Ultsch & Mörchen, 2005)
Differential abundance (DA)	Differential coverage patterns across multiple samples where population changed in abundance over time.	Profile based correlation cut-off.	(iv) Must have multiple samples with population changed in abundance over time, and therefore are associated with cost, computational time, and resources.	(ii), (iii) Strain level resolution can be achieved	(i) Multi-metagenome (Albersten *et al.*, 2013) (ii) MGS Canopy algorithm (Nielsen *et al.*, 2014) (iii) Databionic ESOM tools (Ultsch & Mörchen, 2005)

The methods listed are the main taxonomic-independent binning approaches. In contrast to reference-based approach, referred to as supervised, these features perform binning without knowledge about the genomes in a sample. Properly, these techniques inferred by machine learning cluster sequences based on mutual dissimilarities inherent to the nucleotides sets grouping contigs, or scaffolds into unlabeled bins according to their genetic characteristics (Herath *et al.*, 2017; Sczyrba *et al.*, 2017). Given the feature patterns used in cluster algorithms, the genome binning approach can be divided into three categories: (i) sequence composition-based; (ii) differential abundance-based; (iii) hybrid strategies that combine both sequence composition and differential abundance (Sangwan *et al.,* 2016; Yu *et al.*, 2020; Mallawaarachchi *et al.*, 2021). Sequence composition-based methods are based on the assumption that each taxon has a unique nucleotide composition, so that sequence features from the same genome are more similar than those from different genomes. The commonly used features are the *k*-mer frequencies (typically TNFs), GC content and essential single copy genes. Abundance-based binning methods are based on the assumption that metagenomic sequences belonging to the same genome have parallel abundance in the same sample, and abundance profiles of the sequences belonging to the same species have correlated patterns across multiple samples, which can be used to discriminate closely-related organisms. Thirdly, hybrid strategies employ techniques such as principal component analysis (PCA), probabilistic models and expectation-maximization (EM) algorithms to combine composition-based and abundance-based methods. (iv) Besides these three main categories, additional information may also be taken into consideration: marker genes (*MyCC*, Lin & Liao, 2016), codon usage feature (*BMC3C*, Jiang *et al.*, 2018), taxonomic alignments (*SolidBin*, Pasolli *et al.*, 2019), information rescued from the assembly graph (*GraphBin*, Mallawaarachichi *et al.*, 2020), and paired-end graphs (*METAMVGL*, Zhang *et al.*, 2021). [Figure taken from Sangwan *et al.,* 2016].

Annex 4. List of the biosynthetic gene clusters (BGCs) referenced by antiSMASH

Label	Description	Added	Last updated
acyl_amino_acids	N-acyl amino acid cluster	4.0	4.1
aminocoumarin	Aminocoumarin cluster	<= 3.0	<= 3.0
amglyccycl	Aminoglycoside/aminocyclitol cluster	<= 3.0	<= 3.0
arylpolyene	Aryl polyene cluster	<= 3.0	<= 3.0
betalactone	beta-lactone containing protease inhibitor	5.0	5.0
blactam	β-lactam cluster	<= 3.0	<= 3.0
bottromycin	Bottromycin cluster	<= 3.0	<= 3.0
butyrolactone	Butyrolactone cluster	<= 3.0	<= 3.0
CDPS	tRNA-dependent cyclodipeptide synthases	5.0	5.0
cyanobactin	Cyanobactins like patellamide (AY986476)	<= 3.0	6.0
cyclic-lactone-autoinducer	agrD-like cyclic lactone autoinducer peptides (AF001782)	6.0	6.0
ectoine	Ectoine cluster	<= 3.0	<= 3.0
epipeptide	D-amino-acid containing RiPPs such as yydF (D78193)	6.0	6.0
fatty_acid	Fatty acid cluster (loose strictness, likely from primary metabolism)	<= 3.0	4.2
furan	Furan cluster	<= 3.0	5.0
fungal-RiPP	Fungal RiPP with POP or UstH peptidase types and a modification	5.0	5.0
glycocin	Glycocin cluster	<= 3.0	<= 3.0
guanidinotides	Pheganomycin-style protein ligase-containing cluster	4.0	6.0
halogenated	Cluster containing a halogenase and thus potentially generating a halogenated product	5.0	5.0
hglE-KS	heterocyst glycolipid synthase-like PKS	5.0	5.0
hserlactone	Homoserine lactone cluster	<= 3.0	<= 3.0
indole	Indole cluster	<= 3.0	4.0
LAP	Linear azol(in)e-containing peptides	<= 3.0	6.0
ladderane	Ladderane cluster	<= 3.0	<= 3.0
lantipeptide class I	Class I lanthipeptide clusters like nisin	4.2	6.0
lantipeptide class II	Class II lanthipeptide clusters like mutacin II (U40620)	4.2	6.0
lantipeptide class III	Class III lanthipeptide clusters like labyrinthopeptin (FN178622)	4.2	6.0
lantipeptide class IV	Class IV lanthipeptide clusters like venezuelin (HQ328852)	4.2	6.0
lantipeptide class V	Glycosylated lanthipeptide/linaridin hybrids like MT210103	5.1	6.0
lassopeptide	Lasso peptide cluster	<= 3.0	5.0
linaridin	Linear arid peptide such as cypemycin (HQ148718) and salinipeptin (MG788286)	<= 3.0	<= 3.0
lipolanthine	Lanthipeptide class containing N-terminal fatty acids such as MG673929	5.0	5.0
melanin	Melanin cluster	<= 3.0	<= 3.0
microviridin	Microviridin cluster	<= 3.0	<= 3.0
NAGGN	N-acetylglutaminylglutamine amide	5.0	5.0
NAPAA	non-alpha poly-amino acids like e-Polylysin	6.0	6.0
nrps	Non-ribosomal peptide synthetase cluster	<= 3.0	<= 3.0
nrps-like	NRPS-like fragment	5.0	5.0
nucleoside	Nucleoside cluster	<= 3.0	5.0
oligosaccharide	Oligosaccharide cluster	<= 3.0	<= 3.0
other	Cluster containing a secondary metabolite-related protein that does not fit into any other category	4.0	5.0
PBDE	Polybrominated diphenyl ether cluster	4.1	4.1
phenazine	Phenazine cluster	<= 3.0	<= 3.0
phosphoglycolipid	Phosphoglycolipid cluster	<= 3.0	<= 3.0
phosphonate	Phosphonate cluster	<= 3.0	<= 3.0
PKS-like	Other types of PKS cluster	5.0	5.0
PpyS-KS	PPY-like pyrone cluster	4.2	4.2
proteusin	Proteusin cluster	<= 3.0	<= 3.0
PUFA	Polyunsaturated fatty acid cluster	<= 3.0	<= 3.0
pyrrolidine	Pyrrolidines like described in BGC0001510	6.0	6.0
ranthipeptide	Cys-rich peptides (aka. SCIFF: six Cys in fourty-five) like in CP001581:3481278-3502939	6.0	6.0
RaS-RiPP	Streptide-like thioether-bond RiPPs	5.0	5.0
redox-cofactor	Redox-cofactors such as PQQ (NC_021985:1458906-1494876)	6.0	6.0
resorcinol	Resorcinol cluster	<= 3.0	<= 3.0
RiPP-like	Other unspecified ribosomally synthesised and post-translationally modified peptide product (RiPP) cluster	4.1	6.0
RRE-containing	RRE-element containing cluster	6.0	6.0
saccharide	Saccharide cluster (loose strictness, likely from primary metabolism)	<= 3.0	<= 3.0
sactipeptide	Sactipeptide cluster	<= 3.0	6.0
siderophore	Siderophore cluster	<= 3.0	<= 3.0
spliceotide	RiPPs containing plpX type spliceases (NZ_KB235920:17899-42115)	6.0	6.0
T1PKS	Type I PKS (Polyketide synthase)	<= 3.0	<= 3.0
T2PKS	Type II PKS	<= 3.0	5.0
T3PKS	Type III PKS	<= 3.0	<= 3.0
terpene	Terpene	<= 3.0	4.1
thioamitides	Thioamitide RiPPs as found in JOBF01000011	5.1	6.0
thioamide-NRP	Thioamide-containing non-ribosomal peptide	5.0	5.0
thiopeptide	Thiopeptide cluster	4.2	5.0
transAT-PKS	Trans-AT PKS	<= 3.0	5.0
transAT-PKS-like	Trans-AT PKS fragment, with trans-AT domain not found	<= 5.0	5.0
tropodithietic-acid	Tropodithietic acid cluster	5.0	5.0

Annex 5. Sequencing of the short reads statistics using FastP

Sample	Specie's Name	total reads	reads discarded	duplication rate	mean length	Q30 bases before filtering	Q30 bases after filtering
S14	*P. pacifica*	40,462244 M	5,14 %	3,37 %	138 bp	89,07 %	90,96 %
S15	*P. pacifica*	31,684068 M	5,89 %	2,95 %	135 bp	88,66 %	90,98 %
S16	*P. pacifica*	44,408236 M	2,59 %	3,17 %	119 bp	88,97 %	91,99 %
S17	*P. vitikainenii*	43,764770 M	4,39 %	3,03 %	126 bp	88,62 %	91,04 %
S18	*P. vitikainenii*	43,361814 M	5,03 %	2,97 %	133 bp	87,92 %	90,73 %
S19	*P. vitikainenii*	43,558774 M	4,22 %	2,94 %	131 bp	89,15 %	91,05 %
S20	*P. vitikainenii*	48,445816 M	2,80 %	3,25 %	126 bp	89,29 %	91,54 %
S21	*P. vitikainenii*	39,812964 M	2,68 %	2,89 %	126 bp	89,60 %	91,61 %
S22	*P. vitikainenii*	36,751428 M	3,99 %	2,70 %	139 bp	88,88 %	90,88 %
S23	*P. vitikainenii*	44,681530 M	2,88 %	2,88 %	135 bp	90,34 %	91,80 %
S24	*P. vitikainenii*	40,230744 M	3,23 %	2,78 %	131 bp	89,25 %	91,17 %
S25	*P. vitikainenii*	37,261726 M	4,36 %	3,04 %	117 bp	88,32 %	91,84 %
S26	*P. vitikainenii*	41,352272 M	4,30 %	2,92 %	131 bp	89,70 %	91,75 %
S27	*P. vitikainenii*	43,987270 M	3,47 %	3,11 %	132 bp	89,49 %	91,32 %
S28	*P. vitikainenii*	45,034228 M	2,00 %	2,71 %	129 bp	89,61 %	91,47 %
S29	*P. vitikainenii*	51,612944 M	4,40 %	3,29 %	136 bp	88,73 %	90,54 %
S30	*P. vitikainenii*	44,629874 M	2,62 %	3,60 %	119 bp	89,74 %	92,28 %
S31	*P. vitikainenii*	45,736480 M	4,49 %	3,21 %	126 bp	88,83 %	91,12 %
S32	*P. vitikainenii*	57,417582 M	5,89 %	4,34 %	127 bp	89,09 %	91,66 %
S33	*P. vitikainenii*	49,703902 M	2,98 %	3,00 %	131 bp	90,12 %	91,78 %
S34	*P. vitikainenii*	51,142728 M	3,09 %	3,09 %	134 bp	90,16 %	92,05 %
S35	*P. vitikainenii*	38,731538 M	4,43 %	3,07 %	128 bp	87,86 %	91,13 %
S36	*P. vitikainenii*	54,782730 M	1,92 %	3,08 %	100 bp	88,44 %	92,65 %
S37	*P. vitikainenii*	41,963408 M	4,25 %	3,02 %	135 bp	88,77 %	91,04 %
S38	*P. vitikainenii*	47,947138 M	3,14 %	2,89 %	134 bp	89,46 %	91,60 %
S39	*P. appalachensis*	48,642446 M	3,83 %	2,75 %	137 bp	89,06 %	90,91 %
S40	*P. appalachensis*	49,634348 M	4,37 %	3,01 %	135 bp	88,85 %	91,18 %
S41	*P. appalachensis*	41,133530 M	5,26 %	2,56 %	136 bp	88,53 %	90,48 %
S42	*P. appalachensis*	42,190898 M	3,70 %	2,80 %	138 bp	89,73 %	91,30 %
S43	*P. appalachensis*	43,911730 M	6,44 %	2,87 %	134 bp	88,29 %	91,01 %
S45	*P. appalachensis*	40,761792 M	2,53 %	2,89 %	136 bp	89,87 %	91,44 %
S46	*P. appalachensis*	46,003984 M	3,32 %	2,83 %	136 bp	89,95 %	91,50 %
S47	*P. appalachensis*	41,650414 M	3,14 %	3,03 %	130 bp	89,56 %	91,55 %
S48	*P. appalachensis*	39,018204 M	2,66 %	2,48 %	138 bp	89,67 %	91,03 %
S49	*P. appalachensis*	49,946686 M	3,73 %	3,24 %	128 bp	89,52 %	91,66 %
S50	*P. appalachensis*	50,270758 M	4,16 %	2,90 %	124 bp	89,12 %	91,77 %
S51	*P. appalachensis*	39,833646 M	3,79 %	3,56 %	125 bp	89,40 %	92,34 %
S52	*P. appalachensis*	44,819358 M	2,88 %	2,75 %	138 bp	90,06 %	91,36 %
S54	*P. appalachensis*	24,062750 M	10,71 %	2,79 %	140 bp	86,37 %	89,72 %
S55	*P. appalachensis*	37,723462 M	5,33 %	2,80 %	134 bp	88,87 %	91,01 %
S56	*P. appalachensis*	42,902764 M	2,21 %	2,98 %	131 bp	90,33 %	92,23 %
S57	*P. appalachensis*	42,055932 M	4,58 %	3,02 %	139 bp	89,46 %	91,02 %
S58	*P. appalachensis*	57,940994 M	2,51 %	4,42 %	121 bp	90,66 %	91,73 %
S59	*P. appalachensis*	46,190622 M	2,93 %	3,13 %	124 bp	89,76 %	91,95 %
S60	*P. appalachensis*	45,328958 M	3,24 %	3,53 %	126 bp	90,25 %	92,35 %
S61	*P. asiatica*	41,879048 M	4,52 %	3,27 %	133 bp	88,37 %	90,61 %
S62	*P. asiatica*	42,704714 M	3,38 %	3,12 %	119 bp	88,77 %	91,62 %
S63	*P. asiatica*	36,629996 M	2,79 %	3,45 %	128 bp	90,00 %	91,85 %
S64	*P. mikado*	43,939612 M	4,71 %	3,44 %	133 bp	88,84 %	90,89 %
S65	*P. mikado*	49,244954 M	3,82 %	3,94 %	123 bp	89,41 %	92,09 %
S66	*P. mikado*	37,261390 M	4,32 %	2,49 %	130 bp	89,80 %	91,93 %
S67	*P. gayae*	34,254190 M	7,31 %	3,28 %	140 bp	88,13 %	90,43 %
S68	*P. gayae*	37,650326 M	3,66 %	2,94 %	136 bp	89,76 %	91,41 %
S91	*P. appalachensis*	38,340196 M	6,94 %	2,49 %	132 bp	88,15 %	90,94 %
S93	*P. appalachensis*	40,264300 M	5,76 %	2,64 %	137 bp	88,83 %	90,01 %

Annex 6. Metagenome assembly statistics using QUAST

Sample	Specie's Name	Total length scaffolds.fasta (>0bp)	Total length scaffolds.fasta (>500 bp)	Number of contigs	Number of contigs (≥1000 bp)	Largest contig (bp)	N50 (bp)	N75 (bp)	L50 (bp)	L75 (bp)	GC content (%)	N's per 100kbp
S14	*P. pacifica*	224207994		93951	32272	319522	10380	1554	4078	19911	50.01	16.40
S15	*P. pacifica*	136318557		49120	22550	397564	7998	2461	4186	11832	44.64	10.21
S16	*P. pacifica*	188987589		76866	28866	348366	9133	1680	4653	16905	46.53	7.24
S17	*P. vitikainenii*	151472755		57058	22119	155465	12115	1826	2632	11254	46.67	17.85
S18	*P. vitikainenii*	184084994		73421	28671	560045	10791	1645	3299	16116	50.02	23.05
S19	*P. vitikainenii*	157593914		62398	24377	228177	10363	1713	2998	13572	48.13	16.29
S20	*P. vitikainenii*	139153397		39901	16069	736719	18677	4191	1744	5539	43.90	11.36
S21	*P. vitikainenii*	191033714		111823	37282	205754	3182	958	8507	39574	50.41	16.92
S22	*P. vitikainenii*	179658830		77263	27991	738990	8348	1479	4516	17979	50.27	22.25
S23	*P. vitikainenii*	246632577		141184	44248	269790	3652	959	8959	47000	53.77	23.48
S24	*P. vitikainenii*	183990692		85400	31911	206166	6559	1285	4540	23188	50.37	24.21
S25	*P. vitikainenii*	147235527		101572	40729	273911	1882	917	16734	45487	46.38	10.60
S26	*P. vitikainenii*	212018514		98978	42462	244346	4866	1384	8151	29993	51.94	11.18
S27	*P. vitikainenii*	225685257		125323	44261	291328	3716	1054	10243	41503	53.84	25.79
S28	*P. vitikainenii*	183087131		94979	41032	216946	3702	1219	10073	32868	50.41	13.47
S29	*P. vitikainenii*	192626051		78445	27532	356087	10699	1662	3501	16160	50.15	15.30
S30	*P. vitikainenii*	161857206		65731	25258	172212	8522	1713	4201	14866	47.12	13.17
S31	*P. vitikainenii*	180438588		69452	30709	186459	8655	1822	4214	16497	49.60	18.25
S32	*P. vitikainenii*	218224099		106711	39385	320886	5844	1184	6079	31119	50.02	15.15
S33	*P. vitikainenii*	131355959		55058	19281	178144	9143	1589	3341	11935	43.77	7.67
S34	*P. vitikainenii*	204955084		142178	47199	291289	2135	843	18962	60074	49.97	13.10
S35	*P. vitikainenii*	179356067		72898	28915	556896	9708	1568	3748	16809	49.45	18.81
S36	*P. vitikainenii*	144457707		58001	24163	224052	8064	1731	3723	13850	45.43	15.73
S37	*P. vitikainenii*	159603386		55527	19921	360831	14557	2549	2408	9041	47.02	11.64
S38	*P. vitikainenii*	232358457		96109	34203	331682	10793	1538	4185	20655	53.11	14.57
S39	*P. appalachensis*	168073799		64267	21226	356108	13017	2046	2829	11128	48.11	10.32
S40	*P. appalachensis*	191969001		74761	31333	2216110	12867	1615	2603	16738	48.92	14.12
S41	*P. appalachensis*	126317352		27849	14025	194175	20363	6032	1588	4353	42.10	7.58
S42	*P. appalachensis*	198320990		90301	27306	356760	9057	1301	4416	20117	51.18	15.53
S43	*P. appalachensis*	176743443		75181	26622	313745	12221	1426	2756	16733	47.50	22.24
S45	*P. appalachensis*	147721720		55149	18325	191269	13552	2226	2680	9190	43.05	9.00
S46	*P. appalachensis*	181822517		74484	26761	314654	10539	1644	3393	15767	48.47	15.76
S47	*P. appalachensis*	193683980		90767	33258	131425	6609	1288	4815	24431	48.35	13.67
S48	*P. appalachensis*	155477696		58790	15621	284144	20895	1887	1739	7537	44.53	9.95
S49	*P. appalachensis*	198168018		103964	29559	198911	6844	1014	5016	28945	49.69	16.96
S50	*P. appalachensis*	153605098		74924	21982	173623	8361	1183	3645	18089	45.61	12.31
S51	*P. appalachensis*	243811924		142490	61348	228024	2667	1084	18706	55851	54.54	9.32
S52	*P. appalachensis*	145031804		67753	20584	206978	8314	1265	3691	15762	45.71	15.04
S54	*P. appalachensis*	159541627		81060	37291	171362	3550	1268	8850	28579	49.49	28.12
S55	*P. appalachensis*	165497199		79079	30910	225043	6109	1225	4938	23363	48.78	18.16
S56	*P. appalachensis*	171402044		118942	43406	109838	2105	874	18272	51496	50.29	10.96
S57	*P. appalachensis*	202602631		91152	36880	337564	6765	1365	4969	25055	51.44	29.57
S58	*P. appalachensis*	648055427		118300	35707	166909	2578	826	13294	46906	49.32	7.82
S59	*P. appalachensis*	244730407		140920	50273	227830	3334	1011	13205	49571	53.06	11.93
S60	*P. appalachensis*	238314309		133795	51979	217486	3355	1081	14106	47514	53.87	10.45
S61	*P. asiatica*	196975071		68902	27292	316126	18253	2064	2204	12066	44.64	16.83
S62	*P. asiatica*	174410120		68765	27917	341374	8211	1962	5016	15886	42.83	8.62
S63	*P. asiatica*	138575675		60847	33882	144241	4066	1785	8898	21725	40.88	9.01
S64	*P. mikado*	167071012		78785	27096	182107	7576	1205	4106	20833	49.70	13.94
S65	*P. mikado*	165289026		82360	23434	239754	7760	1077	3954	21022	46.83	12.37
S66	*P. mikado*	120378724		44116	21539	198391	6896	2431	4438	11654	39.46	15.47
S67	*P. gayae*	145930439		48811	16717	519793	25831	2469	1127	6520	42.60	19.23
S68	*P. gayae*	170950990		68111	27742	358920	8869	1769	4110	15632	47.75	13.84
S91	*P. appalachensis*	244628251		120774	45753	223804	5333	1179	6664	36333	52.38	26.49
S93	*P. appalachensis*	166418160		59776	22686	418094	13052	2164	2735	10784	47.70	17.39

Annex 7. Binning statistics of the metagenome-constructed genomes using CONCOCT and Metabat2

Sample	Binning software	Bin	Total length	Nbr scaffolds	N50	GC content	Coverage	CheckM_lineage	CheckM contamination	CheckM completeness	EukCC_lineage	EukCC completeness	EukCC contamination	GTDB_lineage
S14	CONCOCT	55	4128577	173	40460	69,71	12,13	c__Betaproteobacteria	0.45	97.01	NA	0	0	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__BOG-994;s__
S14	CONCOCT	24	3238185	152	34286	71,37	11,72	o__Actinomycetales	1.14	95.65	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Amnibacterium;s__
S14	CONCOCT	16	7810121	250	56928	41,57	113,40	p__Cyanobacteria	0.37	98.78	Eukaryota	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S14	METABAT	7	4482291	422	14123	71,23	10,20	o__Rhizobiales	2.04	93.44	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S14	CONCOCT	0	1664876	877	1927	32,54	4,65	g__Staphylococcus	1.41	59.94	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S14	CONCOCT	8	5524629	875	8418	59,87	8,16	k__Bacteria	4.58	87.99	NA	0	0	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__;s__
S14	METABAT	1	32216727	1821	28221	44,33	21,57	k__Bacteria	16.75	43.54	Eukaryota_Fungi_Ascomycota	80.42	1.64	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__QMWL01;f__;g__;s__
S15	CONCOCT	24	4296180	49	158629	70,19	20,32	o__Rhizobiales	0.73	99.37	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S15	CONCOCT	68	8068875	363	37058	41,52	43,31	p__Cyanobacteria	0.22	98.78	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S15	CONCOCT	16	5702558	1598	4355	50,93	8,45	o__Cytophagales	2.4	90.81	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Spirosoma;s__
S15	METABAT	10	35335918	3023	15339	45,46	25,18	k__Bacteria	14.27	42.94	Eukaryota_Fungi	95.12	2.36	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__
S16	METABAT	12	4064481	44	157153	70,58	20,51	o__Rhizobiales	1.04	98.99	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S16	METABAT	6	3602597	45	127625	71,13	24,21	c__Alphaproteobacteria	1.3	99.08	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S16	METABAT	4	7799941	256	54198	41,51	94,86	p__Cyanobacteria	0.41	98.56	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S16	CONCOCT	45	3144238	405	10172	71,30	11,10	o__Actinomycetales	3.07	96.07	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Amnibacterium;s__
S16	CONCOCT	38	27378399	2190	16894	45,12	14,85	k__Bacteria	11.08	35.25	Eukaryota_Fungi_Ascomycota	80.13	1.85	NA
S17	METABAT	4	6779284	148	74729	41,53	144,45	p__Cyanobacteria	0.52	98.78	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S17	CONCOCT	13	4135497	241	24699	63,25	12,68	f__Bradyrhizobiaceae	1.49	93.82	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S17	CONCOCT	26	4821024	1193	4825	68,91	8,37	o__Actinomycetales	2.96	74.41	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Jatrophihabitans;s__
S17	METABAT	12	29790736	1673	29084	44,87	25,40	k__Bacteria	15.76	43.35	Eukaryota_Fungi	79.46	0.84	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__UBA583;f__;g__;s__
S18	METABAT	17	6859150	153	70820	41,57	65,13	p__Cyanobacteria	0.52	98.56	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S18	CONCOCT	51	3062631	51	122495	68,10	27,56	o__Sphingomonadales	0.7	96.4	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S18	CONCOCT	42	3132965	1165	3126	63,82	6,34	k__Bacteria	5.29	90.13	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S18	METABAT	18	3288345	104	60525	66,33	18,67	o__Sphingomonadales	2.64	98.04	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S18	CONCOCT	47	3677751	360	13833	63,99	9,73	f__Bradyrhizobiaceae	1.91	87.17	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S18	CONCOCT	9	2208224	1068	2262	69,47	5,81	p__Actinobacteria	0.43	46.46	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__70-9;g__VAYN01;s__
S18	CONCOCT	0	30789640	1636	29910	44,09	21,41	k__Bacteria	17.04	44.72	Eukaryota_Fungi	80.81	1.35	NA
S19	METABAT	12	6860845	159	68093	41,59	140,40	p__Cyanobacteria	1.89	97.89	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S19	CONCOCT	12	4027723	1084	4645	65,60	7,50	o__Actinomycetales	1.92	81.3	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Mycobacteriaceae;g__Mycobacterium;s__
S19	METABAT	8	2999938	351	11557	68,18	10,05	o__Sphingomonadales	2.89	91.00	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S19	CONCOCT	28	30702070	1715	29207	44,43	22,51	k__Bacteria	17.67	44.4	Eukaryota_Fungi	82.49	1.18	NA
S20	CONCOCT	42	5771262	40	215498	71,84	30,09	o__Actinomycetales	1.09	96.26	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Iso899;s__
S20	METABAT	11	6988293	153	75659	41,60	161,83	p__Cyanobacteria	0.3	99.00	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S20	METABAT	4	3759203	575	7596	70,39	9,42	o__Rhizobiales	2.06	82.56	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S20	CONCOCT	29	31003329	1673	31074	44,15	29,94	k__Bacteria	17.07	44.72	Eukaryota_Fungi_Basidiomycota	76.74	0	NA
S21	METABAT	6	7993711	273	45256	41,54	61,88	p__Cyanobacteria	0.52	97.89	Eukaryota_Fungi_Microsporidia_Enterocytozoonidae_Enterocytozoon_Enterocytozoon.bieneusi	23.64	10.91	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S21	CONCOCT	56	32264537	2318	18472	44,24	15,89	k__Bacteria	18.40	46.60	Eukaryota_Fungi_Basidiomycota	87.21	1.16	NA
S22	CONCOCT	40	4349958	21	316948	72,55	34,22	p__Actinobacteria	0.43	99.57	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__VAWY01;s__
S22	METABAT	6	8234010	313	42825	41,49	79,54	p__Cyanobacteria	0.67	97.89	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S22	METABAT	11	4594235	375	15963	65,44	10,12	o__Actinomycetales	2.22	97.14	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Mycobacteriaceae;g__Mycobacterium;s__
S22	CONCOCT	3	2952066	566	6845	73,69	8,53	c__Alphaproteobacteria	2.15	89.69	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S22	CONCOCT	49	3958543	2172	1892	68,11	6,36	c__Alphaproteobacteria	45.22	74.05	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas_I;s__
S22	CONCOCT	13	32047357	2185	20949	44,66	15,82	k__Bacteria	15.66	46.44	Eukaryota_Fungi	87.04	1.18	NA
S23	METABAT	12	6827622	148	69602	41,60	68,96	p__Cyanobacteria	0.52	97.44	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S23	CONCOCT	21	4232573	65	112584	70,39	17,78	o__Actinomycetales	1.99	93.48	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Jatrophihabitans;s__
S23	CONCOCT	27	34473921	1872	27671	43,65	20,29	k__Bacteria	17.77	45.42	Eukaryota_Fungi	87.88	2.02	NA
S24	METABAT	4	7049992	167	65017	41,58	127,07	p__Cyanobacteria	0.52	99.00	Eukaryota	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S24	CONCOCT	12	4158611	576	9087	65,49	8,79	o__Actinomycetales	0.88	87.48	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Mycobacteriaceae;g__Mycobacterium;s__
S24	METABAT	10	4300943	781	6220	63,33	7,66	f__Bradyrhizobiaceae	4.94	75.95	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S24	CONCOCT	37	34202997	1875	27867	43,78	22,05	k__Bacteria	17.60	45.92	Eukaryota_Fungi	87.54	2.36	NA
S25	METABAT	12	2753093	62	87744	32,59	13,75	g__Staphylococcus	0.19	99.51	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S25	METABAT	6	4106527	139	48588	70,67	22,11	o__Rhizobiales	0.73	96.98	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S25	METABAT	7	3928954	395	12917	66,75	11,48	o__Sphingomonadales	1.91	93.76	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S25	METABAT	5	8168787	302	44872	41,52	108,47	p__Cyanobacteria	0.74	98.78	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S25	METABAT	1	6337909	1043	7048	60,57	8,88	k__Bacteria	3.40	78.83	NA	0	0	d__Bacteria;p__Armatimonadota;c__BOG-944;o__Capsulimonadales;f__;g__;s__
S26	METABAT	12	5528648	99	90458	72,00	18,83	o__Actinomycetales	1.44	93.46	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Iso899;s__
S26	CONCOCT	52	4489644	79	101416	63,80	31,86	k__Bacteria	3.49	98.51	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S26	CONCOCT	49	8867345	332	44913	41,54	146,85	p__Cyanobacteria	0.77	99.00	Eukaryota_Fungi_Microsporidia	22.92	10.42	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S26	CONCOCT	17	3354799	90	81652	65,68	33,53	c__Alphaproteobacteria	0.00	98.74	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas_I;s__
S26	CONCOCT	50	4138024	261	25349	70,55	17,89	o__Rhizobiales	1.36	91.93	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S26	CONCOCT	35	4418649	128	60948	70,22	16,74	o__Rhizobiales	2.19	95.94	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S26	METABAT	20	28104387	3016	11360	45,54	12,14	k__Bacteria	11.79	42.49	Eukaryota_Fungi	86.87	3.2	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__
S27	CONCOCT	46	8071476	121	115978	66,60	17,40	o__Actinomycetales	2.82	99.03	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Mycobacteriaceae;g__Mycobacterium;s__
S27	METABAT	15	4145909	80	91002	70,56	17,76	o__Rhizobiales	0.73	97.35	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S27	METABAT	4	7292649	179	60413	41,82	93,77	p__Cyanobacteria	0.22	99.56	Eukaryota_Euglenozoa_Diplonemea_Diplonemidae_Diplonema_Diplonema.papillatum	11.11	5.56	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S27	CONCOCT	34	31480839	2638	14998	44,69	12,34	k__Bacteria	12.81	43.50	Eukaryota_Fungi	89.06	2.02	NA
S28	METABAT	4	3877840	191	31215	67,94	15,55	o__Rhizobiales	1.58	98.49	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylopilaceae;g__Hansschlegelia;s__
S28	METABAT	7	7315872	264	45835	41,58	295,04	p__Cyanobacteria	0.52	98.56	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S28	METABAT	5	4243894	145	45074	63,82	15,91	k__Bacteria	3.41	98.35	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S28	CONCOCT	0	4351008	599	9196	69,12	11,89	o__Actinomycetales	3.10	84.30	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Jatrophihabitans;s__
S28	METABAT	2	2697459	499	5770	43,80	9,00	p__Bacteroidetes	0.63	75.71	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Sphingobacteriales;f__Sphingobacteriaceae;g__Mucilaginibacter;s__
S28	METABAT	13	31727555	3324	11970	45,22	12,26	k__Bacteria	13.09	43.15	Eukaryota_Fungi	92.26	2.36	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__

Annex 7. (suit)

Sample	Binning software	Bin	Total length	Nbr scaffolds	N50	GC content	Coverage	CheckM_lineage	CheckM contamination	CheckM completeness	EukCC_lineage	EukCC completeness	EukCC contamination	GTDB_lineage
S29	METABAT	2	3398952	50	141092	65,41	31,49	c__Alphaproteobacteria	0.00	97.33	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas_I;s__
S29	CONCOCT	28	4415186	242	31730	70,12	14,06	c__Alphaproteobacteria	5.41	96.10	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S29	METABAT	12	6790194	159	77166	41,63	208,55	p__Cyanobacteria	0.41	98.33	Eukaryota_Rhodophyta_Bangiophyceae_Cyanidiales_Cyanidiaceae	18.87	4.85	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S29	CONCOCT	44	5682139	889	8785	71,46	9,74	o__Actinomycetales	1.17	91.11	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Iso899;s__
S29	METABAT	6	4975074	229	36427	70,70	18,31	o__Rhizobiales	1.25	97.14	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S29	CONCOCT	8	34153626	1896	26902	44,04	25,12	k__Bacteria	15.38	44.72	Eukaryota_Fungi	82.61	0.0	NA
S30	METABAT	7	7074277	196	59630	41,50	153,50	p__Cyanobacteria	2.07	98.78	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S30	METABAT	5	4333791	114	65478	63,90	40,01	k__Bacteria	3.04	98.51	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S30	METABAT	13	4375212	428	12889	71,33	11,51	o__Rhizobiales	2.28	95.29	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S30	CONCOCT	3	4404618	581	10541	63,53	13,91	k__Bacteria	3.58	92.59	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S30	METABAT	11	2967087	168	27222	68,33	16,80	o__Sphingomonadales	2.53	96.45	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S31	CONCOCT	10	2537900	567	6113	32,58	6,21	g__Staphylococcus	1.34	91.79	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S31	METABAT	13	8091091	188	56053	41,64	135,87	p__Cyanobacteria	0.52	98.11	Eukaryota_Fungi_Microsporidia	22.92	8.33	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S31	METABAT	11	3531938	300	16653	70,54	13,64	o__Actinomycetales	3.23	92.98	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Amnibacterium;s__
S31	CONCOCT	4	4333937	1989	2412	72,09	7,02	o__Actinomycetales	7.52	71.81	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Iso899;s__
S31	CONCOCT	14	3687984	947	4963	72,73	10,38	k__Bacteria	6.24	76.10	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__VAWY01;s__
S31	CONCOCT	6	34708526	2227	22344	44,40	26,53	k__Bacteria	17.54	47.65	Eukaryota_Fungi	90.07	1.85	NA
S32	CONCOCT	10	2751354	34	107861	32,59	17,43	g__Staphylococcus	0.08	99.51	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S32	METABAT	10	4401330	146	58433	68,62	20,36	o__Rhizobiales	1.43	98.64	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S32	METABAT	1	4252831	102	78979	63,32	24,35	f__Bradyrhizobiaceae	1.08	97.91	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S32	METABAT	13	6777123	197	49945	41,61	171,89	p__Cyanobacteria	0.22	98.78	Eukaryota_Rhodophyta	14.04	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S32	METABAT	21	3134981	61	92302	68,58	38,09	o__Sphingomonadales	0.97	98.41	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S32	CONCOCT	67	3292674	1172	3319	65,31	8,71	o__Actinomycetales	3.11	81.20	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Jatrophihabitantaceae;g__Iso899;s__
S32	CONCOCT	14	2824049	587	5531	63,68	9,22	k__Bacteria	5.74	73.41	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S32	METABAT	2	3611827	529	8274	68,33	10,57	c__Alphaproteobacteria	11.16	80.73	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__B12;s__
S32	CONCOCT	30	6769875	683	12775	67,77	11,27	k__Bacteria	56.03	94.12	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S32	METABAT	9	29842736	1823	23447	45,33	30,47	k__Bacteria	12.17	41.57	Eukaryota_Fungi_Basidiomycota_Agaricomycetes_Agaricales	78.64	2.54	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__
S33	METABAT	3	6780332	201	50525	41,66	296,50	p__Cyanobacteria	0.30	97.44	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S33	CONCOCT	52	34223005	2378	19237	44,09	17,68	k__Bacteria	16.55	48.17	Eukaryota_Fungi	92.09	2.02	NA
S34	METABAT	2	6725415	152	70018	41,54	218,25	p__Cyanobacteria	0.30	98.78	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S34	CONCOCT	15	6519916	894	9713	46,38	9,14	p__Bacteroidetes	1.82	88.15	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__;s__
S34	CONCOCT	30	4992614	806	7359	65,35	9,50	o__Actinomycetales	1.46	89.25	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Mycobacteriaceae;g__Mycobacterium;s__
S34	CONCOCT	43	24740178	3656	7993	46,01	10,78	k__Bacteria	9.87	39.78	Eukaryota_Fungi	80.81	4.71	NA
S35	CONCOCT	26	3071079	41	129352	68,25	31,30	o__Sphingomonadales	0.82	99.48	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S35	METABAT	6	6868135	184	69455	41,61	123,41	p__Cyanobacteria	0.41	98.33	Eukaryota_Rhodophyta_Bangiophyceae_Cyanidiales_Cyanidiaceae	18.6	5.93	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S35	METABAT	1	4258421	153	41404	64,08	16,57	k__Bacteria	5.11	97.72	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S35	METABAT	8	3955454	280	19874	71,54	11,31	o__Rhizobiales	2.28	89.99	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S35	METABAT	4	5039419	462	15212	70,83	10,49	o__Rhizobiales	3.03	94.65	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S35	METABAT	3	29244387	2138	19388	45,01	19,06	k__Bacteria	15.34	41.79	Eukaryota_Fungi	78.26	0.0	d__Archaea;p__Asgardarchaeota;c__Heimdallarchaeia;o__LC-3;f__;g__;s__
S36	METABAT	2	4481788	87	88476	65,65	21,42	o__Rhizobiales	0.94	98.98	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S37	METABAT	2	4009928	120	68017	68,21	16,74	c__Alphaproteobacteria	0.36	98.89	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__B12;s__
S37	METABAT	8	6966227	161	77827	41,51	184,15	p__Cyanobacteria	2.07	97.00	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S37	METABAT	12	4234542	363	15613	69,15	11,25	o__Rhizobiales	3.55	91.01	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S37	CONCOCT	17	4088489	1423	3447	69,07	7,93	o__Rhizobiales	11.91	83.43	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S37	CONCOCT	34	31177023	1895	25789	44,21	19,67	k__Bacteria	17.54	45.58	Eukaryota_Fungi	85.02	1.18	NA
S38	METABAT	16	6292746	142	65176	54,39	16,83	o__Cytophagales	1.49	99.40	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__;s__
S38	METABAT	11	3085794	42	191265	68,22	30,10	o__Sphingomonadales	0.37	98.52	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S38	METABAT	17	7117908	167	79095	41,53	194,18	p__Cyanobacteria	2.07	98.56	Eukaryota_Rhodophyta	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__Nostoc sp002949735
S38	METABAT	27	3173635	57	93867	73,48	27,08	c__Alphaproteobacteria	1.30	99.15	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S38	METABAT	25	4443437	166	52856	70,67	24,78	o__Rhizobiales	1.86	97.52	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S38	CONCOCT	40	4307193	486	11911	64,23	10,20	k__Bacteria	1.35	92.58	Eukaryota_Fungi_Ascomycota	8.7	4.35	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S38	METABAT	12	5112790	186	42378	63,50	25,25	k__Bacteria	1.38	97.16	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S38	CONCOCT	49	2885624	1346	2348	43,75	6,07	p__Bacteroidetes	1.83	69.60	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Sphingobacteriales;f__Sphingobacteriaceae;g__Mucilaginibacter;s__
S38	METABAT	20	3662746	198	28857	70,74	14,11	o__Rhizobiales	6.71	87.24	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S38	CONCOCT	6	31515205	1916	24006	44,25	20,06	k__Bacteria	17.46	45.06	Eukaryota_Fungi	85.35	1.18	NA
S39	CONCOCT	9	4830618	76	125602	67,92	15,21	c__Betaproteobacteria	1.07	99.79	NA	0	0	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__BOG-994;s__
S39	CONCOCT	45	4254796	246	25831	70,45	10,85	o__Rhizobiales	0.89	97.66	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S39	METABAT	8	8191194	431	29719	41,90	156,13	p__Cyanobacteria	0.41	94.89	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S39	METABAT	5	3100673	417	9611	66,62	8,93	o__Actinomycetales	0.95	88.96	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__;g__;s__
S39	CONCOCT	38	7732371	1898	5157	65,17	8,28	k__Bacteria	5.11	86.57	Eukaryota_Fungi_Ascomycota_Dothideomycetes	11.92	2.4	d__Bacteria;p__Planctomycetota;c__Planctomycetes;o__Gemmatales;f__Gemmataceae;g__Gemmata;s__
S39	CONCOCT	35	33305024	1712	32344	43,97	24,11	k__Bacteria	13.88	43.86	Eukaryota_Fungi	84.34	1.85	NA
S40	CONCOCT	35	3210651	2	2216110	69,84	33,97	o__Actinomycetales	1.01	99.49	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Schumannella;s__
S40	METABAT	20	4397949	84	97490	64,30	14,72	f__Bradyrhizobiaceae	2.04	97.69	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S40	CONCOCT	20	3506109	55	140968	67,60	31,79	o__Sphingomonadales	1.05	99.14	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S40	METABAT	18	7299480	135	79090	41,94	149,85	p__Cyanobacteria	0.44	98.00	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S40	CONCOCT	6	6162028	2853	2383	71,73	6,42	k__Bacteria	5.22	71.23	Eukaryota_Euglenozoa	8.33	0.0	d__Bacteria;p__Planctomycetota;c__Planctomycetes;o__Gemmatales;f__Gemmataceae;g__PALSA-1347;s__
S40	CONCOCT	47	33862254	1522	38341	43,52	31,89	k__Bacteria	13.30	40.06	Eukaryota_Fungi_Microsporidia_Nosematidae_Nosema	70.81	16.75	NA
S41	METABAT	8	7618213	374	29897	41,93	110,97	p__Cyanobacteria	0.41	94.22	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S41	METABAT	3	3877493	954	4062	69,82	7,09	o__Rhodospirillales	19.15	74.94	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S41	CONCOCT	13	32653394	1421	40742	43,88	31,75	k__Bacteria	14.03	41.30	Eukaryota_Fungi	83.0	1.52	NA
S42	METABAT	1	4260159	106	68389	71,52	14,68	o__Actinomycetales	3.11	98.23	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Amnibacterium;s__
S42	CONCOCT	24	6061969	119	109882	46,85	16,14	k__Bacteria	7.74	95.73	NA	0	0	d__Bacteria;p__Acidobacteriota;c__Blastocatellia;o__Pyrinomonadales;f__Pyrinomonadaceae;g__GCA-002050365;s__
S42	METABAT	13	4232119	151	47484	64,28	15,22	o__Rhizobiales	1.58	96.30	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__RH-AL1;s__
S42	METABAT	5	7475994	397	27916	41,95	162,50	p__Cyanobacteria	0.48	93.56	Eukaryota_Fungi_Microsporidia	18.75	10.42	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S42	CONCOCT	32	4280548	1001	5355	62,83	7,40	f__Bradyrhizobiaceae	1.17	85.23	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S42	CONCOCT	29	3830377	259	19217	70,66	10,76	o__Rhizobiales	4.19	86.07	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S42	CONCOCT	18	6636472	853	9137	71,44	8,95	o__Rhizobiales	47.86	90.07	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S42	CONCOCT	1	34601494	2176	22410	44,03	17,41	k__Bacteria	17.38	44.72	Eukaryota_Fungi	89.9	2.19	NA
S43	METABAT	10	8524251	226	67551	41,62	61,91	p__Cyanobacteria	0.44	98.78	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__

Annex 7. (suit)

Sample	Binning software	Bin	Total length	Nbr scaffolds	N50	GC content	Coverage	CheckM_lineage	CheckM contamination	CheckM completeness	EukCC_lineage	EukCC completeness	EukCC contamination	GTDB_lineage
S45	METABAT	5	7988487	390	29870	41,85	164,91	p__Cyanobacteria	0.49	93.22	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S45	CONCOCT	41	2336055	801	3468	70,60	7,67	o__Actinomycetales	1.03	80.39	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Frankiaceae;g__;s__
S46	METABAT	3	8493280	219	67584	41,62	125,49	p__Cyanobacteria	0.44	98.33	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S46	METABAT	5	3901112	545	8717	70,51	9,60	o__Rhodospirillales	2.50	90.02	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Geminicoccales;f__Geminicoccaceae;g__Arboricoccus;s__
S46	CONCOCT	17	3978188	1277	3814	59,00	7,11	k__Bacteria	5.32	92.61	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__UBA10450;g__AV80;s__
S46	CONCOCT	19	28779396	1706	27922	44,41	19,73	k__Bacteria	14.33	38.37	Eukaryota_Fungi	75.08	1.01	NA
S47	METABAT	6	7275875	410	25925	42,05	92,43	p__Cyanobacteria	0.37	92.66	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S47	CONCOCT	41	32201444	1899	24285	43,91	19,26	k__Bacteria	18.24	41.97	Eukaryota_Fungi	86.7	2.02	NA
S48	METABAT	13	3321630	70	71671	69,95	14,52	o__Actinomycetales	1.73	98.99	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Schumannella;s__
S48	METABAT	5	8164301	136	106111	41,44	94,91	p__Cyanobacteria	0.52	99.78	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S48	CONCOCT	29	32910754	1672	33502	43,60	25,15	k__Bacteria	14.45	42.13	Eukaryota_Fungi	78.96	2.02	NA
S49	METABAT	14	4036561	588	8118	68,99	9,35	o__Rhodospirillales	1.59	87.34	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Acetobacterales;f__Acetobacteraceae;g__BOG-908;s__
S49	CONCOCT	25	8973099	321	57012	41,41	41,35	p__Cyanobacteria	0.44	99.22	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S49	METABAT	11	4129224	488	10312	64,36	10,13	f__Bradyrhizobiaceae	4.95	89.56	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S49	CONCOCT	16	32677983	1967	22968	43,92	18,34	k__Bacteria	18.21	42.13	Eukaryota	72.73	4.13	NA
S50	METABAT	1	4708577	189	43801	60,93	16,06	k__Bacteria	0.86	97.53	NA	0	0	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__Bryocella;s__
S50	CONCOCT	48	7000833	2867	2766	41,94	6,63	p__Cyanobacteria	3.11	79.94	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S50	CONCOCT	56	33454205	1923	23702	43,89	17,39	k__Bacteria	19.05	45.24	Eukaryota_Fungi	88.05	2.36	NA
S51	METABAT	22	3216227	167	28543	71,17	13,60	o__Actinomycetales	1.83	96.58	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Microbacteriaceae;g__Amnibacterium;s__
S51	CONCOCT	53	6605199	395	23644	53,65	13,62	o__Cytophagales	1.49	99.33	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__;s__
S51	CONCOCT	66	2492374	677	4839	32,56	6,16	g__Staphylococcus	2.98	88.88	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S51	METABAT	10	7928416	304	50267	41,69	108,98	p__Cyanobacteria	0.33	98.88	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S51	METABAT	20	3107108	110	46108	65,30	15,24	o__Rhizobiales	1.83	90.53	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__RH-AL1;s__
S51	METABAT	15	4273384	322	19177	72,17	18,88	o__Rhizobiales	1.61	87.82	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S52	METABAT	7	6937367	169	64940	41,87	246,90	p__Cyanobacteria	0.56	97.33	Eukaryota_Fungi_Microsporidia	18.75	12.5	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S52	CONCOCT	28	32184334	1968	22304	44,33	15,67	k__Bacteria	18.88	42.29	Eukaryota_Fungi	89.73	2.02	NA
S54	METABAT	17	8520291	332	44111	41,62	77,78	p__Cyanobacteria	2.19	96.88	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S54	CONCOCT	36	2526014	593	5876	32,63	6,04	g__Staphylococcus	2.04	91.97	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S54	CONCOCT	37	32318880	2569	16044	45,23	13,70	k__Bacteria	11.82	40.41	Eukaryota_Fungi	93.27	2.69	NA
S55	CONCOCT	37	2707361	205	23462	32,57	8,18	g__Staphylococcus	1.58	98.12	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S55	METABAT	6	7047296	193	57966	41,83	122,37	p__Cyanobacteria	1.22	97.11	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S55	CONCOCT	3	35739047	2129	23583	44,53	24,57	k__Bacteria	18.56	43.18	Eukaryota_Fungi	91.75	2.02	NA
S56	CONCOCT	30	8640538	419	35915	41,77	209,66	p__Cyanobacteria	1.00	97.33	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S56	METABAT	8	3338342	593	6022	62,19	8,19	k__Bacteria	3.07	74.59	NA	0	0	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__EB88;s__
S56	CONCOCT	77	4830048	843	7112	70,46	10,13	o__Rhodospirillales	2.70	92.23	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Acetobacterales;f__Acetobacteraceae;g__;s__
S56	METABAT	10	3674545	588	7062	56,85	8,79	k__Bacteria	1.96	86.99	NA	0	0	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__Bryocella;s__
S56	CONCOCT	27	2915780	998	3405	69,17	8,14	c__Alphaproteobacteria	2.60	79.31	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__B12;s__
S56	METABAT	7	24241453	4361	6113	45,77	9,24	k__Bacteria	8.86	38.25	Eukaryota_Fungi	78.62	4.38	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__
S57	CONCOCT	40	7477125	152	76960	50,26	14,32	o__Cytophagales	1.19	98.81	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__;s__
S57	METABAT	8	7889065	287	47851	41,71	132,77	p__Cyanobacteria	1.30	96.66	Eukaryota	21.74	4.35	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S57	CONCOCT	58	4250499	378	16208	70,95	13,25	o__Rhizobiales	10.36	86.98	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S57	CONCOCT	5	2846486	1770	1615	69,24	5,15	c__Betaproteobacteria	15.53	59.66	NA	0	0	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__;s__
S57	CONCOCT	53	35488106	1792	29945	43,71	22,88	k__Bacteria	16.81	44.20	Eukaryota_Fungi	88.22	1.68	NA
S58	METABAT	4	7410359	237	47831	41,80	110,88	p__Cyanobacteria	0.56	97.22	Eukaryota_Rhodophyta_Bangiophyceae_Cyanidiales_Cyanidiaceae	15.79	1.75	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S58	CONCOCT	6	3311089	624	6282	66,46	14,26	c__Alphaproteobacteria	0.73	82.55	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__Palsa-881;s__
S58	CONCOCT	33	2662947	971	3242	57,61	8,32	k__Bacteria	1.21	80.17	NA	0	0	d__Bacteria;p__Eremiobacterota;c__Eremiobacteria;o__UBP12;f__UBA5184;g__;s__
S58	CONCOCT	45	27162844	3328	9272	44,91	10,83	k__Bacteria	10.59	42.18	Eukaryota_Fungi	83.33	4.71	NA
S59	METABAT	8	4297867	95	75881	70,58	26,87	o__Rhizobiales	1.61	96.33	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S59	METABAT	4	3696882	110	73917	72,21	24,87	c__Alphaproteobacteria	2.27	99.08	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S59	CONCOCT	73	4742638	234	36980	63,53	32,11	k__Bacteria	3.60	97.16	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S59	METABAT	1	7356769	247	46023	41,89	105,46	p__Cyanobacteria	0.63	97.33	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S59	CONCOCT	60	5744687	1873	3715	53,72	10,76	o__Cytophagales	2.52	87.18	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__;s__
S59	METABAT	9	4863878	250	29557	71,06	18,34	o__Rhizobiales	4.89	95.82	Eukaryota	8.7	0.0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S59	METABAT	20	4156433	122	51772	70,60	20,77	k__Bacteria	5.80	94.33	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S59	METABAT	2	28371717	3144	11091	45,43	12,63	k__Bacteria	10.51	39.04	Eukaryota_Fungi	86.87	3.7	d__Archaea;p__Nanoarchaeota;c__Nanoarchaeia;o__PEZQ01;f__;g__;s__
S60	METABAT	1	7143502	214	64475	41,83	131,06	p__Cyanobacteria	0.56	96.77	Eukaryota_Euglenozoa	11.11	2.78	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S60	CONCOCT	20	3348545	626	6771	73,19	9,34	o__Actinomycetales	0.99	89.04	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__;g__;s__
S60	CONCOCT	18	3637206	610	8150	63,85	9,64	k__Bacteria	1.39	91.00	NA	0	0	d__Bacteria;p__Eremiobacterota;c__Eremiobacteria;o__UBP12;f__UBA5184;g__Palsa-1515;s__
S60	METABAT	14	2464543	428	6334	73,68	9,84	c__Alphaproteobacteria	3.19	77.72	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S60	METABAT	10	3321341	197	24635	71,01	16,09	o__Rhizobiales	4.75	82.39	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S60	CONCOCT	27	4479677	729	7845	63,91	11,28	k__Bacteria	16.33	86.01	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S60	CONCOCT	41	27378950	2564	13180	45,30	14,17	k__Bacteria	13.20	40.59	Eukaryota_Fungi	84.18	3.54	NA
S61	METABAT	13	4404908	151	48824	70,31	11,97	o__Rhizobiales	1.42	98.55	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S61	METABAT	14	7618852	233	59126	41,50	39,56	p__Cyanobacteria	0.44	98.88	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S61	CONCOCT	1	3733102	1193	3814	60,79	6,65	k__Bacteria	0.93	82.32	Eukaryota_Rhodophyta	8.77	0.0	d__Bacteria;p__Armatimonadota;c__BOG-944;o__Capsulimonadales;f__Capsulimonadaceae;g__;s__
S61	CONCOCT	18	3675951	1830	2206	57,74	5,73	k__Bacteria	0.00	72.91	NA	0	0	d__Bacteria;p__Armatimonadota;c__BOG-944;o__Capsulimonadales;f__;g__;s__
S61	CONCOCT	25	32346103	1487	35545	43,35	25,12	k__Bacteria	18.09	42.16	Eukaryota_Fungi	81.48	1.85	NA

Annex 7. (suit)

Sample	Binning software	Bin	Total length	Nbr scaffolds	N50	GC content	Coverage	CheckM_lineage	CheckM contamination	CheckM completeness	EukCC_lineage	EukCC completeness	EukCC contamination	GTDB_lineage
S62	CONCOCT	26	4475243	569	11847	69,45	17,52	k__Bacteria	2.88	93.06	NA	0	0	d__Bacteria;p__Eremiobacterota;c__Eremiobacteria;o__UBP12;f__UBA5184;g__Palsa-1478;s__
S62	METABAT	13	7443291	131	113067	41,59	99,18	p__Cyanobacteria	0.52	98.88	Eukaryota_Fungi_Basidiomycota	8.14	1.16	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S62	CONCOCT	27	26796026	2400	14149	45,00	13,45	k__Bacteria	10.11	37.84	Eukaryota_Fungi	81.48	2.19	NA
S63	METABAT	1	7627861	405	28036	42,00	199,08	p__Cyanobacteria	0.22	95.11	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S63	CONCOCT	25	1791728	872	2171	32,63	4,98	g__Staphylococcus	2.50	70.68	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S63	CONCOCT	21	4003537	1266	3836	70,37	8,41	o__Rhizobiales	4.47	88.05	Eukaryota_Euglenozoa	5.56	0.0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S63	CONCOCT	20	24584019	3917	7234	45,75	11,55	k__Archaea	15.84	53.13	Eukaryota_Fungi	82.15	4.88	NA
S64	CONCOCT	29	3189031	304	13564	73,42	12,44	c__Alphaproteobacteria	2.11	94.13	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__PMMR1;s__
S64	METABAT	6	8033583	372	32342	41,93	177,76	p__Cyanobacteria	0.78	96.00	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S64	METABAT	9	2892076	283	13543	69,81	11,43	c__Alphaproteobacteria	2.19	87.65	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__;s__
S64	METABAT	4	4589271	437	13480	71,24	13,03	o__Rhizobiales	4.39	88.49	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Methylobacterium;s__
S64	CONCOCT	52	31401344	1722	29428	44,06	22,26	k__Bacteria	13.10	42.68	Eukaryota_Fungi_Ascomycota	82.49	1.52	NA
S65	METABAT	1	8087613	204	91693	41,54	163,61	p__Cyanobacteria	0.52	98.89	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S65	CONCOCT	32	2467001	753	4183	32,59	6,00	g__Staphylococcus	2.40	92.02	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S65	METABAT	7	3850488	427	11688	69,59	11,82	o__Rhizobiales	2.04	89.45	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S65	CONCOCT	45	27655956	1978	19562	44,61	17,12	k__Bacteria	14.22	45.27	Eukaryota_Fungi	78.45	2.02	NA
S66	METABAT	4	7602722	372	30145	42,01	75,82	p__Cyanobacteria	1.33	96.22	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S66	CONCOCT	60	28672574	2613	13182	44,87	12,29	k__Bacteria	14.22	40.45	Eukaryota_Fungi	86.87	2.69	NA
S67	METABAT	8	7928955	437	26155	41,99	83,55	p__Cyanobacteria	0.52	96.67	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S67	CONCOCT	14	2332614	790	3595	32,62	5,61	g__Staphylococcus	2.79	84.62	NA	0	0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Staphylococcales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus aureus
S67	METABAT	15	1349753	338	3967	70,06	6,65	o__Actinomycetales	0.54	46.22	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Frankiaceae;g__;s__
S67	CONCOCT	7	5220518	2885	2548	71,68	6,02	k__Bacteria	70.72	97.96	NA	0	0	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__;f__;g__;s__
S67	CONCOCT	24	31925513	1779	37382	42,58	21,77	k__Bacteria	11.32	36.65	Eukaryota_Fungi	73.4	1.18	NA
S68	METABAT	14	4724380	154	50893	71,34	14,03	o__Rhizobiales	2.66	98.12	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S68	CONCOCT	14	4393880	68	114835	53,45	13,92	p__Bacteroidetes	0.49	99.01	NA	0	0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Flavisolibacter;s__
S68	CONCOCT	10	8805062	621	23270	41,99	155,05	p__Cyanobacteria	0.63	98.56	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S68	CONCOCT	16	4828328	717	9231	61,61	9,91	k__Bacteria	4.26	94.91	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S68	CONCOCT	1	3109126	683	5652	64,85	9,27	o__Sphingomonadales	1.25	87.47	Eukaryota_Fungi_Ascomycota	8.42	0.88	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Novosphingobium;s__
S68	CONCOCT	33	3213571	1020	3764	60,33	7,12	k__Bacteria	3.50	84.42	NA	0	0	NA
S68	METABAT	16	6088143	1416	4298	69,61	7,44	o__Rhodospirillales	42.74	83.14	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Acetobacterales;f__Acetobacteraceae;g__;s__
S68	CONCOCT	38	29625949	2101	21129	44,53	15,41	k__Bacteria	11.71	45.06	Eukaryota_Fungi	80.98	2.53	NA
S91	METABAT	20	4560102	113	67251	63,74	29,08	k__Bacteria	4.39	98.51	NA	0	0	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__;g__;s__
S91	METABAT	21	7613700	412	26247	41,99	49,90	p__Cyanobacteria	0.48	93.77	Eukaryota_Fungi_Microsporidia	18.75	12.5	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S91	METABAT	15	4573095	258	27491	69,98	12,17	o__Rhizobiales	1.29	97.35	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S91	METABAT	13	4024354	105	64730	67,11	18,93	o__Rhizobiales	1.16	98.85	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Lichenihabitans;s__
S91	METABAT	9	3136451	300	13732	69,23	13,70	o__Sphingomonadales	4.97	89.46	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
S91	CONCOCT	47	4709885	894	6449	71,29	8,05	o__Rhizobiales	17.08	80.39	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Beijerinckiaceae;g__Enterovirga;s__
S91	CONCOCT	30	34099184	2089	23327	43,69	16,41	k__Bacteria	15.66	44.91	Eukaryota_Evosea_Eumycetozoa	70.91	6.91	NA
S93	METABAT	6	4280139	42	215796	62,92	26,95	f__Bradyrhizobiaceae	0.63	97.27	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Tardiphaga;s__
S93	CONCOCT	26	8970127	463	31444	41,95	81,53	p__Cyanobacteria	0.56	98.89	NA	0	0	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Cyanobacteriales;f__Nostocaceae;g__Nostoc;s__
S93	METABAT	8	3401167	267	17683	66,17	10,15	k__Bacteria	6.54	82.37	NA	0	0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__;s__
S93	CONCOCT	9	31189627	1513	34902	44,19	25,47	k__Bacteria	13.49	38.14	Eukaryota_Fungi	80.3	1.68	NA

Annex 8. Class rules classification for biosynthetic gene clusters evaluated by BiG-SCAPE

[https://github.com/medema-group/BiG-SCAPE]

BiG-SCAPE classes

By default, BiG-SCAPE will try to separate clusters into eight classes. This is done based on the `product` annotation from antiSMASH (see here). New labels introduced in antiSMASH 5 and 6 are annotated with a superscript.

The rules followed are currently:

antiSMASH annotation	BiG-SCAPE class
`t1pks`, `T1PKS`[5]	PKS I
`transatpks`, `t2pks`, `t3pks`, `otherks`, `hglks`, `transAT-PKS`[5], `transAT-PKS-like`[5], `T2PKS`[5], `T3PKS`[5], `PKS-like`[5], `hglE-KS`[5] and combinations of these with {`t1pks`, `T1PKS`[5]} or themselves	PKS other
`nrps`, `NRPS`[5], `NRPS-like`[5], `thioamide-NRP`[5], `NAPAA`[6]	NRPS
`lantipeptide`, `thiopeptide`, `bacteriocin`, `linaridin`, `cyanobactin`, `glycocin`, `LAP`, `lassopeptide`, `sactipeptide`, `bottromycin`, `head_to_tail`, `microcin`, `microviridin`, `proteusin`, `guanidinotides`, `RiPP-like`, `lanthipeptide`[5], `lipolanthine`[5], `RaS-RiPP`[5], `fungal-RiPP`[5], `thioamitides`[5].1, `lanthipeptide-class-i`[6], `lanthipeptide-class-ii`[6], `lanthipeptide-class-iii`[6], `lanthipeptide-class-iv`[6], `lanthipeptide-class-v`[6], `ranthipeptide`[6], `redox-cofactor`[6], `RRE-containing`[6], `epipeptide`[6], `cyclic-lactone-autoinducer`[6], `spliceotide`[6] and combinations of these	RiPPs
`amglyccycl`, `oligosaccharide`, `cf_saccharide`, `saccharide`[5] and combinations of these	Saccharides
`terpene`	Terpene
any of {PKS I} + any of {NRPS}	PKS/NRPS Hybrids
`acyl_amino_acids`, `arylpolyene`, `aminocoumarin`, `ectoine`, `butyrolactone`, `nucleoside`, `melanin`, `phosphoglycolipid`, `phenazine`, `phosphonate`, `other`, `cf_putative`, `resorcinol`, `indole`, `ladderane`, `PUFA`, `furan`, `hserlactone`, `fused`, `cf_fatty_acid`, `siderophore`, `blactam`, `fatty_acid`[5], `PpyS-KS`[5], `CDPS`[5], `betalactone`[5], `PBDE`[5], `tropodithietic-acid`[5], `NAGGN`[5], `halogenated`[5], `pyrrolidine`[6] and any combined annotation	Others
*	< mix >

Hybrids

If the `hybrids` mode is enabled, some clusters could be analyzed in different classes (if those are valid classes):

If cluster is the PKS-NRP_Hybrids BiG-SCAPE class: the cluster will also be put in the NRPS class and one of the PKS classes (PKS I or PKS other). If the cluster contains the `t1pks` annotation, it will *always* be put on the PKS I class.

If the cluster is classified as Others and BiG-SCAPE detects that it's because of a multiple annotation (e.g. `terpene-t1pks`), BiG-SCAPE will also put the cluster in every different individual class.

Glossary

Accession number: An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence (or other) entry.

Active site: The amino acid residues at the catalytic site of an enzyme. These residues provide the binding and activation energy needed to place the substrate into its transition state and bridge the energy barrier of the reaction undergoing catalysis

Adapters: The oligonucleotides bound to the 5'and 3'end of each DNA fragment in a sequencing library. The adapters are complementary to the lawn of oligonucleotides present on the surface of Illumina sequencing flow cells.

Algorithm: A series of steps defining a procedure or formula for solving a problem, that can be coded into a programming language and executed. Bioinformatics algorithms typically are used to process, store, analyze, visualize and make predictions from biological data.

Alignment: The result of a comparison of two or more gene or protein sequences in order to determine their degree of base or amino acid similarity. Sequence alignments are used to determine the similarity, homology, function or other degree of relatedness between two or more genes or gene products.

Analogy: Reasoning by which the function of a novel gene or protein sequence may be deduced from comparisons with other gene or protein sequences of known function. Identifying analogous or homologous genes via similarity searching and alignment is one of the chief uses of Bioinformatics. (See also alignment, similarity search.)

Annotation: A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

Assembly: Compilation of overlapping sequences from one or more related genes that have been clustered together based on their degree of sequence identity or similarity. Sequence assembly may be used to piece together "shotgun" sequencing fragments (see shotgun sequencing) based upon overlapping restriction enzyme digests, or may be used to identify and index novel genes from "single-pass" cDNA sequencing efforts.

Base pair: A pair of nitrogenous bases (a purine and a pyrimidine), held together by hydrogen bonds, that form the core of DNA and RNA i.e. the A:T, G:C and A:U interactions.

Bioinformatics: The field of endeavor that relates to the collection, organization and analysis of large amounts of biological data using networks of computers and databases (usually with reference to the genome project and DNA sequence information)

cDNA (complementary DNA): A DNA strand copied from mRNA using reverse transcriptase. A cDNA library represents all of the expressed DNA in a cell.

cDNA library: A set of DNA fragments prepared from the total mRNA obtained from a selected cell, tissue or organism.

Chimeric clone: A cloning artifact created by a foreign gene being inserted into a vector in an incorrect orientation resulting in the expression of a protein consisting of a fusion of two different gene products.

Cluster: The grouping of similar objects in a multidimensional space. Clustering is used for constructing new features which are abstractions of the existing features of those objects. The quality of the clustering depends crucially on the distance metric in the space. In bioinformatics, clustering is performed on sequences, high-throughput expression and other experimental data. Clusters of partial or complete gene sequences can be used to identify the complete (contiguous) sequence and to better identify its function. Clustering expression data enables the researcher to discern patterns of co-regulation in groups of genes.

Coding regions (CDS): The portion of a genomic sequence bounded by start and stop codons that identifies the sequence of the protein being coded for by a particular gene.

Codon: A sequence of three adjacent nucleotides that designates a specific amino acid or start/stop site for transcription.

Consensus sequence: A single sequence delineated from an alignment of multiple constituent sequences that represents a "best fit" for all those sequences. A "voting" or other selection procedure is used to determine which residue (nucleotide or amino acid) is placed at a given position in the event that not all of the constituent sequences have the identical residue at that position.

Contig: A length of contiguous sequence assembled from partial, overlapping sequences, generated from a "shotgun" sequencing project. Contigs are typically created computationally, by comparing the overlapping ends of several sequencing reads generated by restriction enzyme digestion of a segment of genomic DNA. The creation of contigs in the presence of sequencing errors, ambiguities and the presence of repeats is one of the most computationally challenging aspects of the role of Bioinformatics in genome analysis.

Coverage level: The average number of sequenced bases that align to each base of the reference DNA.

Data Cleaning: A process whereby automated or semi-automated algorithms are used to process experimental data, including noise, experimental errors and other artifacts, in order to generate and store high-quality data for use in subsequent analysis. Data cleaning is typically required in high-throughput sequencing where compression or other experimental artifacts limit the amount of sequence data generated from each sequencing run or "read."

Data Mining: The ability to query very large databases in order to satisfy a hypothesis ("top-down" data mining); or to interrogate a database in order to generate new hypotheses based on rigorous statistical correlations ("bottom-up" data mining).

Database: Any file system by which data gets stored following a logical process. (see also relational database)

Dendrogram: A graphical procedure for representing the output of a hierarchical clustering method. A dendrogram is strictly defined as a binary tree with a distinguished root, that has all the data items at its leaves. Conventionally, all the leaves are shown at the same level of the drawing. The ordering of the leaves is arbitrary, as is their horizontal position. The heights of the internal nodes may be arbitrary, or may be related to the metric information used to form the clustering.

DNA (deoxyribonucleic acid): The chemical that forms the basis of the genetic material in virtually all organisms. DNA is composed of the four nitrogenous bases Adenine, Cytosine, Guanine, and Thymine, which are covalently bonded to a backbone of deoxyribose - phosphate to form a DNA strand. Two complementary strands (where all Gs pair with Cs and As with Ts) form a double helical structure which is held together by hydrogen bonding between the cognate bases.

DNA sequencing: The technique in which the specific sequence of bases forming a particular DNA region is deciphered.

Domain (protein): A region of special biological interest within a single protein sequence. However, a domain may also be defined as a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein sequences that accomplishes a specific function. A domain class is a group of domains that share a common set of well-defined properties or characteristics.

Enzyme: A class of proteins that are capable of catalyzing chemical reactions (the making or breaking of chemical bonds). They do so by orienting their substrates into a suitable geometry in a particular location (the active site) where electrophilic or nucleophilic amino acid residues can participate in the reaction. Enzymes are protein catalyst that speeds up chemical reactions that would otherwise be prohibitively slow under physiological conditions.

Epigenomics: The study of complex expression networks or linkages both spatially (within the body) and temporally (at different times in development).

Eukaryote: A cell or organism with a distinct membrane-bound nucleus as well as specialized membrane-based organelles (see also prokaryote).

Exon: The region of DNA within a gene that codes for a polypeptide chain or domain. Typically a mature protein is composed of several domains coded by different exons within a single gene.

Expression (gene or protein): A measure of the presence, amount, and time-course of one or more gene products in a particular cell or tissue. Expression studies are typically performed at the RNA (mRNA) or protein level in order to determine the number, type, and level of genes that may be up-regulated or down-regulated during a cellular process, in response to an external stimulus, or in sickness or disease. Gene chips and proteomics now allow the study of expression profiles of sets of genes or even entire genomes.

Gaps (affine gaps): A gap is defined as any maximal, consecutive run of spaces in a single string of a given alignment. Gaps help create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment. The idea is to take in account the number of continuous gaps and not only the number of spaces when calculating an alignment. Affine gaps contain a component for gap insertion and a component for gap extension, where the extension penalty is usually much lower than the insertion penalty. This mimics biological reality as multiple gaps would imply multiple mutations, but a single mutation can lead to a long gap quite easily.

GenBank: Data bank of genetic sequences operated by a division of the National Institutes of Health.

Gene: Classically, a unit of inheritance. In practice, a gene is a segment of DNA on a chromosome that encodes a protein and all the regulatory sequences (promoter) required to control expression of that protein.

Gene expression: The conversion of information from gene to protein via transcription and translation.

Gene families: Subsets of genes containing homologous sequences which usually correlate with a common function.

Gene library: A collection of cloned DNA fragments created by restriction endonuclease digestion that represent part or all of an organism's genome.

Gene product: The product, either RNA or protein, that results from expression of a gene. The amount of gene product reflects the activity of the gene.

Genetic code: The mapping of all possible codons into the 20 amino acids including the start and stop codons.

Genetic marker: Any gene that can be readily recognized by its phenotypic effect, and which can be used as a marker for a cell, chromosome, or individual carrying that gene. Also, any detectable polymorphism used to identify a specific gene.

Genome: The complete genetic content of an organism.

Genomic DNA (sequence): DNA sequence typically obtained from mammalian or other higher-order species, which includes both intron and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter, and enhancer sequences.

Genomics: The analysis of the entire genome of a chosen organism.

Genotype: Strictly, all of the genes possessed by an individual. In practice, the particular alleles present in a specific genetic locus.

Hidden Markov model (HMM): A joint statistical model for an ordered sequence of variables. The result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"), where the Markov chain has discrete variables which select the "state" of the HMM at each step. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-

dependent scoring parameters such as gap penalties, thus more accurately modeling the consequences of evolutionary events on sequence families.

High-throughput screening: The method by which very large numbers of compounds are screened against a putative drug target in either cell-free or whole-cell assays. Typically, these screenings are carried out in 96 well plates using automated, robotic station based technologies or in higher- density array ("chip") formats.

Holobiont : an assemblage of a host and the many other species living in or around it, which together form a discrete ecological unit.

Homology: (strict) Two or more biological species, systems or molecules that share a common evolutionary ancestor. (general) Two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein), that they register along their lengths. Sequence "homology" searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage and possible function of the query gene.

in silico (biology): (Lit. computer mediated). The use of computers to simulate, process, or analyze a biological experiment.

Introns: Nucleotide sequences found in the structural genes of eukaryotes that are non-coding and interrupt the sequences containing information that codes for polypeptide chains. Intron sequences are spliced out of their RNA transcripts before maturation and protein synthesis. (cf. Exons)

Iteration: A series of steps in an algorithm whereby the processing of data is performed repetitively until the result exceeds a particular threshold. Iteration is often used in multiple sequence alignments whereby each set of pairwise alignments are compared with every other, starting with the most similar pairs and progressing to the least similar, until there are no longer any sequence-pairs remaining to be aligned.

Library: A large collection of compounds, peptides, cDNAs or genes which may be screened in order to isolate cognate molecules.

Locus: The specific position occupied by a gene on a chromosome. At a given locus, any one of the variant forms of a gene may be present. The variants are said to be alleles of that gene.

Markov chain: Any multivariate probability density whose independence diagram is a chain. The variables are ordered, and each variable "depends" only on its neighbors in the sense of being conditionally independent of the others. Markov chains are an integral component of hidden Markov models.

Metabolomics: This term describes the analytical approaches used to determine the metabolite profile(s) in any given strain or single tissue. The resulting census of all metabolites present in any given strain or single tissue is called the *metabolome*. Most commonly used platforms to characterize the metabolome include nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) linked to a liquid chromatography separation system.

Metagenome: The collection of genomes and genes from the members of a microbiota. This collection is obtained through shotgun sequencing of DNA extracted from a sample (metagenomics) followed by assembly or mapping to a reference database followed by annotation. Metataxonomic analysis, because it relies on the amplification and sequencing of taxonomic marker genes, is not metagenomics. Metagenomics is the process used to characterize the metagenome, from which information on the potential function of the microbiota can be gained. Metagenomics was first used by Handelsman et al.; however, it was in the context of what the authors called functional metagenomics, an approach where random fragments of environmental DNA are cloned into a suitable vector for maintenance in a surrogate host for functional screening, looking for gain of function in the surrogate host.

Microarray: A 2D array, typically on a glass, filter, or silicon wafer, upon which genes or gene fragments are deposited or synthesized in a predetermined spatial order allowing them to be made available as probes in a high-throughput, parallel manner.

Modeling: In bioinformatics, modeling usually refers to molecular modeling, a process whereby the three-dimensional architecture of biological molecules is interpreted (or predicted), visually represented, and manipulated in order to determine their molecular properties. (general) A series of mathematical equations or procedures which simulate a real-life process, given a set of assumptions, boundary parameters, and initial conditions.

Motif: A conserved element of a protein sequence alignment that usually correlates with a particular function. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be predictive of any subsequent occurrence of such a structural/functional region in any other novel protein sequence.

Multigene family: A set of genes derived by duplication of an ancestral gene, followed by independent mutational events resulting in a series of independent genes either clustered together on a chromosome or dispersed throughout the genome.

Multiple (sequence) alignment: A Multiple Alignment of k sequences is a rectangular array, consisting of characters taken from the alphabet A, that satisfies the following conditions: There are exactly k rows; ignoring the gap character, row number i is exactly the sequence sI; and each column contains at least one character different from "-". In practice multiple sequence alignments include a cost/weight function, that defines the penalty for the insertion of gaps (the "-" character) and weights identities and conservative substitutions accordingly. Multiple alignment algorithms attempt to create the optimal alignment defined as the one with the lowest cost/weight score.

Nucleotide: A nucleic acid unit composed of a five carbon sugar joined to a phosphate group and a nitrogen base.

Oligonucleotide: A short molecule consisting of several linked nucleotides (typically between 10 and 60) covalently attached by phosphodiester bonds.

Open reading frame (ORF): Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.

Operator: A segment of DNA that interacts with the products of regulatory genes and facilitates the transcription of one or more structural genes.

Operon: A unit of transcription consisting of one or more structural genes, an operator, and a promoter.

Ortholog: Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.)

Pattern: Molecular biological patterns usually occur at the level of the characters making up the gene or protein sequence. A pattern language must be defined in order to apply different criteria to different positions of a sequence. In order to have position-specific comparison done by a computer, a pattern-matching algorithm must allow alternative residues at a given position, repetitions of a residue, exclusion of alternative residues, weighting, and ideally, combinatorial representation.

Pathways: Bioinformatics strives to define representations of key biological data types, algorithms and inference procedures, including sequences, structures, biological pathways and reactions. Representing and computing with biological pathways requires ontologies for representing pathway knowledge; User interfaces to these databases; Physicochemical properties of enzymes and their substrates in pathways; And pathway analysis of whole genomes including identifying common patterns across species and species differences.

Paralog: Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

Parameters: Parameters are user-selectable values, typically experimentally determined, that govern the boundaries of an algorithm or program. For instance, selection of the appropriate input parameters governs the success of a search algorithm. Some of the most common search parameters in bioinformatics tools include the stringency of an alignment search tool, and the weights (penalties) provided for mismatches and gaps.

Peptide: A short stretch of amino acids each covalently coupled by a peptide (amide) bond.

Pharmacogenomics: The use of (DNA-based) genotyping in order to target pharmaceutical agents to specific patient populations. Genetic differences are known to affect responses to many types of drug therapy, and pharmacogenomics analysis serves to customize the use of pharmaceuticals for specific subgroups of patients. The rationale for this approach is that observed gene expression differences may correlate with, and explain, the differences in side effects and efficacy to drugs in humans.

Phenotype: Any observable feature of an organism that is the result of one or more genes.

Phylum: The segmentation of the animal kingdom into about 30 major groups collectively known as phyla. The members of each phylum share the same basic structure and organization. For instance, fish, birds, and human beings belong to one phylum - the Chordata - because all have spinal cords.

Primer: A short oligonucleotide that provides a free 3í hydroxyl for DNA or RNA synthesis by the appropriate polymerase (DNA polymerase or RNA polymerase).

Profile: Sequence profiles are usually derived from multiple alignments of sequences with a known relationship, and consist of tables of position-specific scores and gap-penalties. Each position in the profile contains scores for all of the possible amino acids, as well as one penalty score for opening and one for continuing a gap at the specified position. Attempts have been made to further improve the sensitivity of the profile by refining the procedures to construct a profile starting from a given multiple alignment. Other representations for sequence domains or motifs do not necessarily require the presence of a correct and complete multiple alignment, such as hidden Markov models.

Prokaryote: An organism or cell that lacks a membrane-bounded nucleus. Bacteria and blue-green algae are the only surviving prokaryotes (cf. Eukaryote).

Promoter (site): A promoter site is defined by its recognition by eukaryotic RNA polymerase II; its activity in a higher eukaryote; by experimentally evidence, or homology and sufficient similarity to an experimentally defined promoter; and by observed biological function.

Protein families: Sets of proteins that share a common evolutionary origin reflected by their relatedness in function which is usually reflected by similarities in sequence, or in primary, secondary or tertiary structure. Subsets of proteins with related structure and function.

Proteomics: The study of the proteome. Typically, the cataloging of all the expressed proteins in a particular cell or tissue type, obtained by identifying the proteins from cell extracts using a combination of 2D gel electrophoresis and mass spectrometry. The large scale analysis of the protein composition and function. (cf genomics)

Query (sequence): A DNA, RNA of protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced.

Read: next generation sequencing uses sophisticated instruments to determine the nucleotide sequence of a DNA or RNA sample. In general terms, a sequence “read” refers to the data string of A,T, C, and G bases corresponding to the sample DNA or RNA. With Illumina technology, millions of reads are generated in a single sequencing run.

Reading frame: A sequence of codons beginning with an initiation codon and ending with a termination codon, typically of at least 150 bases (50 amino acids) coding for a polypeptide or protein chain (see ORF and URF).

Reference genome: A reference genome is a fully sequenced and assembled genome that acts as a scaffold

against which new sequence reads are aligned and compared.

Relational Database: A database that follows E. F. Coddi's 11 rules, a series of mathematical and logical steps for the organization and systemization of data into a software system that allows easy retrieval, updating, and expansion. An RDBMS stores data in a database consisting of one or more tables of rows and columns. The rows correspond to a record (tuple); the columns correspond to attributes (fields) in the record. In an RDBMS, a view, defined as a subset of the database that is the result of the evaluation of a query, is a table. RDBMSs use Structured Query Language (SQL) for data definition, data management, and data access and retrieval. Relational and object-relational databases are used extensively in bioinformatics to store sequence and other biological data.

Repeats (repeat sequences): Repeat sequences and approximate repeats occur throughout the DNA of higher organisms (mammals). For example, the Alu sequences of length about 300 characters, appear hundreds of thousands of times in Human DNA with about 87% homology to a consensus Alu string. Some short substrings such as TATA-boxes, poly-A and (TG)* also appear more often than by chance. Repeat sequences may also occur within genes, as mutations or alterations to those genes. Repetitive sequences, especially mobile elements, have many applications in genetic research. DNA transposons and retroposons are routinely used for insertional mutagenesis, gene mapping, gene tagging, and gene transfer in several model systems.

Repetitive elements: Repetitive elements provide important clues about chromosome dynamics, evolutionary forces, and mechanisms for exchange of genetic information between organisms The most ubiquitous class of repetitive elements in the DNA sequence in primate genomes is the Alu family of interspersed repeats which have arisen in the last 65 million years of evolution Alu repeats belong to a class of sequences defined as short interspersed elements (SINEs). Approximately 500,000 Alu SINEs exist within the human genome, representing about 5% of the genome by mass.

Replication: The synthesis of an informationally identical macromolecule (e.g. DNA) from a template molecule.

Ribonucleic acid (RNA): A category of nucleic acids in which the component sugar is ribose and consisting of the four nucleotides Thymidine, Uracil, Guanine, and Adenine. The three types of RNA are messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

Sensitivity: Sensitivity of bioinformatics similarity search algorithms centers around two areas: First, how well can the method detect biologically meaningful relationships between two related sequences in the presence of mutations and sequencing errors; Secondly how does the heuristic nature of the algorithm affect the probability that a matching sequence will not be detected. At the user's discretion, the speed of most similarity search programs can be sacrificed in exchange for greater sensitivity - with an emphasis on detecting lower scoring matches.

Shotgun cloning: The cloning of an entire gene segment or genome by generating a random set of fragments using restriction endonucleases to create a gene library that can be subsequently mapped and sequenced to reconstruct the entire genome.

Similarity (homology) search: Given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

Supervised learning: data mining implementation that uses a set of independent attributes to predict the value of a dependent attribute or target.

Unidentified reading frame (URF): An open reading frame encoding a protein of undefined function.

Unsupervised learning: data mining implementation that does not distinguish between dependent and independent attributes.

Maud Debras, Mémoire de recherche, 2022

Sous la direction du Professeur Denis Baurain et du Docteur Luc Cornet

Master en Sciences Bioinformatiques, Finalité approfondie, Université de Liège, 7 Place du XX Août, 4000 Liège

Analysis of secondary metabolite biosynthesis gene clusters in lichen metagenomes

Lichens are organisms resulting from the symbiotic association between a fungus and a photosynthetic partner, a green alga or a cyanobacterium. The unique chemical diversity that results from this way of life reveals numerous compounds with various therapeutic interests.

To exploit this potential, I use a recent method of genome reconstruction for six species of Peltigera lichens and highlight the capabilities of automated bioinformatics tools in the context of this metagenomic study. After obtaining the genome assemblies, the genetic motifs involved in the biosynthesis of secondary metabolites of cyanobionts were investigated with automated bioinformatics tools.

In a first step, this study demonstrated the validity of the metagenomic assembly protocol to reconstruct the genome of cyanobionts of the genus Nostoc from Peltigera species collected in the wild. In a second step, the investigation of the Nostoc genomes revealed a remarkably high biosynthetic potential and promise for the discovery of new natural products. The gene clusters mainly detected were associated with the NRPS (non-ribosomal peptide synthetases) and PKS (polyketide synthases) enzyme synthesis pathways. On the other hand, the majority of the genetic biosynthetic pathways detected in this study had no biological description to date. These observations highlighted the interest in automated methods in the search for new metabolomic profiles. Finally, the identification of particularly rich secondary metabolite synthesis potentials revealed by cyanobionts from P.appalacchensis, P.asiatica, P.borenquensis, and P.mikado species, suggested a genetic heritage.

Analyse des clusters de gènes de biosynthèse des métabolites secondaires dans les métagénomes des lichens

Les lichens sont des organismes resultants de l'association symbiotique entre un champignon et un partenaire photosynthétique, une algue verte ou une cyanobactérie. La diversité chimique unique qui résulte de ce mode de vie révèle de nombreux composés munis d'intérêts thérapeutiques variés.

Afin de valoriser ce potentiel, j'utilise dans ce travail une méthode récente de reconstruction de génomes pour six espèces de lichens *Peltigera*, et mets en évidence les capacités des outils bioinformatiques automatisés dans le contexte de cette étude metagénomique. Après obtention des assemblage de génomes, les motifs génétiques impliqués dans la biosynthèse des métabolites secondaires des cyanobiontes ont été investigués avec des outils bioinformatiques automatisés.

Dans un premier temps, cette étude a permis de démontrer la validité du protocole d'assemblage métagénomique pour reconstruire le génome du cyanobionte du genre *Nostoc* issus d'espèces *Peltigera* prélevés en milieu naturel. Dans un second temps, l'investigation des génomes de *Nostoc* a révélé un potentiel biosynthétiques remarquablement élevé et prometteur pour la découverte de nouveaux produits naturels. Les groupes de gènes principalement détectés étaient associés aux voies de synthèses des enzymes NRPS (non-ribosomal peptides synthetases) et PKS (polyketides synthases). D'autre part, la majorité des voies génétiques de biosynthèse détectées lors de cette étude ne bénéficiaient d'aucune description biologique à ce jour. Ces observations ont mis en évidence l'intérêt des méthodes automatisées dans la recherche de nouveaux profils métabolomiques. Finalement, la mise en évidence de potentiels de synthèse de métabolites secondaires particulièrement riches révélés par les cyanobiontes issus des espèces *P.appalacchensis*, *P.asiatica*, *P.borenquensis* et *P.mikado,* suggérait un patrimoine génétique intéressant pour un de futures investigations.

Keywords: Lichen; Peltigera; Nostoc; Metagenomics; Metabolites