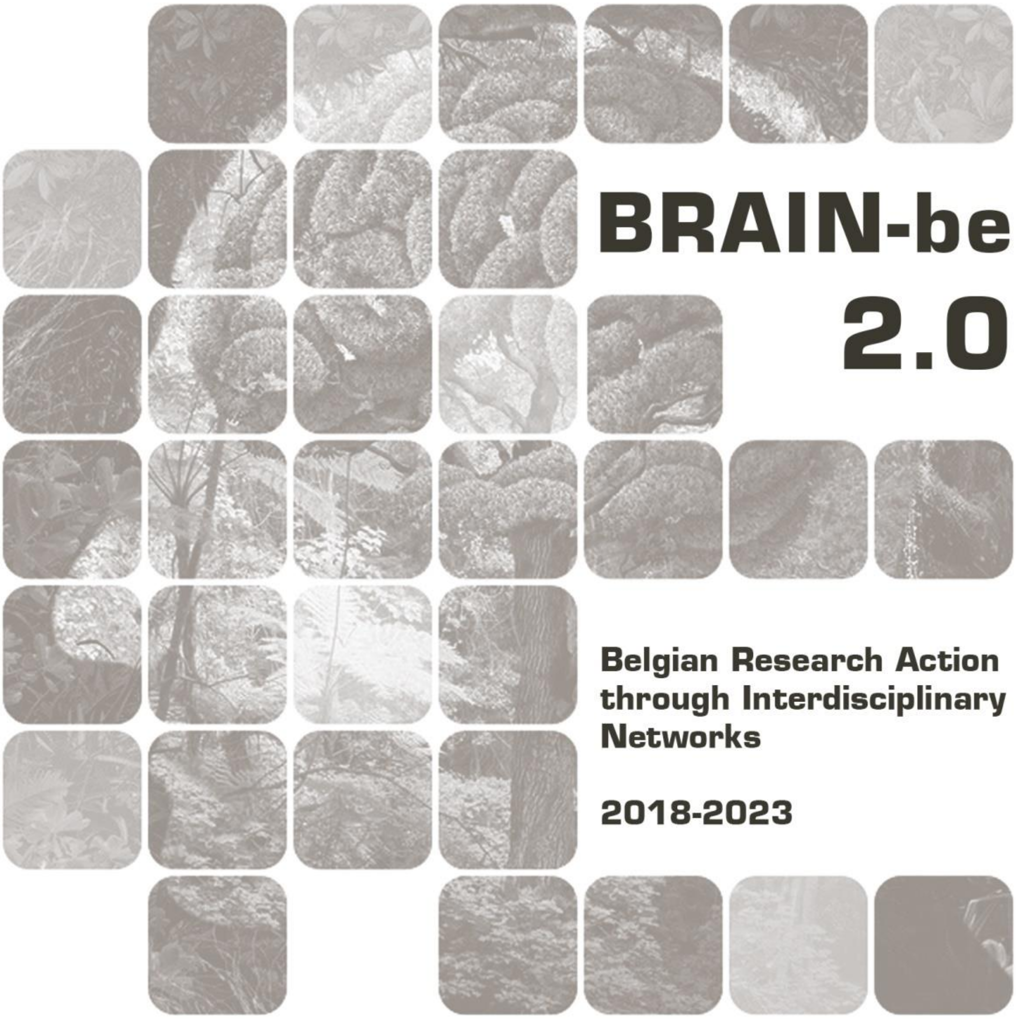


BRAIN-be
2.0

Belgian Research Action
through Interdisciplinary
Networks

2018-2023

BCCM GEN-ERA

BCCM collections in the genomic era

Pierre Becker (Sciensano) – Luc Cornet (Sciensano) – Elizabet D’hooge (Sciensano) – Ilse Cleenwerck (UGent) – Oren Tzfadia (ITG) – Leen Rigouts (ITG) – Wim Mulders (ITG) – Heide-Marie Daniel (UCLouvain) – Annick Wilmotte (ULiège) – Denis Baurain (ULiège)

Pillar 2: Heritage science

NETWORK PROJECT

BCCM GEN-ERA

BCCM collections in the genomic era

Contract - B2/191/P2/BCCM GEN-ERA

FINAL REPORT

PROMOTORS:	Pierre Becker (Sciensano) Denis Baurain (ULiège) Stephan Declerck (UCLouvain) Leen Rigouts (ITG) Peter Vandamme (UGent) Annick Wilmotte (ULiège)
AUTHORS:	Pierre Becker (Sciensano) Luc Cornet (Sciensano) Elizabet D’hooge (Sciensano) Ilse Cleenwerck (UGent) Oren Tzfadia (ITG) Leen Rigouts (ITG) Wim Mulders (ITG) Heide-Marie Daniel (UCLouvain) Annick Wilmotte (ULiège) Denis Baurain (ULiège)

Published in 2022 by the Belgian Science Policy Office
WTCIII
Simon Bolivarlaan 30 bus 7
Boulevard Simon Bolivar 30 bte 7
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Georges JAMART
Tel: +32 (0)2 238 36 90

Neither the Belgian Science Policy Office nor any person acting on behalf of the Belgian Science Policy Office is responsible for the use which might be made of the following information. The authors are responsible for the content.

Pierre Becker, Luc Cornet, Elizabet D'hooge, Ilse Cleenwerck, Oren Tzfadia, Leen Rigouts, Wim Mulders, Heide-Marie Daniel, Annick Wilmotte, Denis Baurain. ***BCCM collections in the genomic era***. Final Report. Brussels: Belgian Science Policy Office 2022 – 40 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT

Context

The Belgian Coordinated Collections of Microorganisms (BCCM) is a unique initiative of the Belgian Federal Government and its Science Policy Office (Belspo). BCCM is a Biological Resource Centre (BRC) that preserves and provides microbial and genetic resources to support life sciences and the biotechnology sector in the field of fundamental and applied research.

Prokaryotic and eukaryotic microorganisms represent most of the biodiversity present on Earth and can be found in every environment capable of supporting life. They play a major role in a huge range of functions, from ecosystems to humans, and contribute to countless fBRCapplications. Nevertheless, the current knowledge in microbiology barely represents the tip of the iceberg and microbiological research still needs to tackle the considerable task of finding and understanding the yet undiscovered microbial capabilities. Culture collections play an essential role in this endeavour by isolating, cultivating, identifying, preserving and distributing the cultivable diversity.

Since its creation in 1983, BCCM developed and maintained international leadership among European BRCs through the implementation of a number of features including a website, an ISO 9001 certification, an online catalogue, a laboratory information management system and its recognition as international depositary authority. The study, valorisation and documentation of the microbial resources also need to stay up to date with the current developments in microbiology and modern technologies to analyse microorganisms in an efficient manner. Nowadays, microbial research is greatly facilitated by new approaches in genomics including whole genome sequencing. The latter provides the entire genetic information of an organism and is increasingly requested in many disciplines. Acquiring know-how in genomics is thus crucial for BCCM in order to remain a major BRC, for future national and international collaborations and to answer upcoming research questions.

Objectives

The first objective of the BCCM GEN-ERA project was to implement expertise in genomics in the BCCM collections for which the real challenge was the handling and analysis of genomic big data. This required the installation of specific bioinformatics structures (hardware) and software for which the BCCM scientists had to be trained to ensure a long-term implementation. The focus was primarily on whole genome sequencing because the determination of the entire gene repertoire of microorganisms supports the leading expertise of the collections in the field of taxonomy and phylogeny while allowing potential functional analyses. Moreover, offering strains having their whole genome sequenced is necessary to meet the needs of BRCs users and is thus an added value for the visibility and attractiveness of BCCM.

The BCCM GEN-ERA project also aimed at answering specific research questions covering the microbial diversity of the BCCM collections (i.e., bacteria, mycobacteria, cyanobacteria,

yeasts and moulds) and more specifically on selected microorganisms having a societal impact (i.e., associated with pollinating insects, causing pathologies or producing bioactive compounds). The project implicated five out of the seven BCCM collections and was a collaboration with the laboratory of Eukaryotic Phylogenomics of the University of Liège that brought the necessary expertise in bioinformatics and (phylo)genomics.

Conclusions

Two different bioinformatics infrastructures were envisaged for the handling and analysis of the sequencing data, namely Galaxy and Nextflow. Both were tested and compared for their respective performance, appropriateness, user-friendliness and compliance with the FAIR principles (i.e., Findable, Accessible, Interoperable, Reusable).

Galaxy is a bioinformatics web platform whose objective is notably to make genomic analyses available to all researchers, even with little computing skills, thanks to a user-friendly graphical user interface. The installation of a private "BCCM" Galaxy was tested but experienced several security gaps that complicated the deployment. Therefore, it appeared that a system administrator would be needed to maintain the infrastructure. These difficulties also limited the interoperability and reuse of the bioinformatics workflows. Moreover, some mandatory programs required for modern bioinformatics practices were not available. For these reasons, Nextflow, designed to perform bioinformatics using command lines instead of a graphical user interface, was preferred. In total, 14 Nextflow workflows, sustained by 11 Singularity containers were implemented. They cover the common genomics needs for microbial taxonomy and metabolic modelling of microbial collections like BCCM. They can be used on prokaryotes and small eukaryotes in a completely reproducible manner. The workflows are provided to the users as a program that can be run with a single command line, increasing the reproducibility of the analyses. This "GEN-ERA toolbox" was made freely available from the GitHub repository https://github.com/Lcornet/GENERA which also includes a large documentation for the users. Nextflow thus fulfilled most of the criteria for a long-term and FAIR utilization of the bioinformatics infrastructure at BCCM. The only disadvantage, as compared to Galaxy, is the user-friendliness. Working with command lines is indeed less intuitive and necessitated dedicated trainings, but could be reduced to a minimum thanks to the Singularity containers.

The Nextflow infrastructure implemented at BCCM in collaboration with the University of Liège was used to investigate case studies for which pending research questions could be addressed by genomic analyses. In particular, we investigated fungal pathogens causing skin infections, non-tuberculosis mycobacteria, bacteria and yeasts from the gut of pollinating insects (bees and bumblebees) as well as cyanobacteria displaying biological activities. These analyses provided breakthroughs in their respective fields and opened new perspectives for future researches.

The BCCM GEN-ERA project established expertise in genomics at BCCM by setting a bioinformatics framework, by bringing genomic tools and by developing genomics skills of the BCCM scientists. This investment was thus performed in view of a long-term implementation

allowing genomics to become a continuous activity within the BCCM consortium. In this respect, an essential milestone was achieved with the BCCM GEN-ERA GitHub repository which can be regarded as an open gateway to use, reuse and learn to use bioinformatic tools for genomic analyses. It was designed to offer online and free access to analysis programs for the handling of genomic data from prokaryotic and eukaryotic microorganisms. It is a web portal that centralizes software for notably genome assembly, genome annotation, phylogenomics or metabolic modelling of genomes. It works as a common platform that can be used by all BCCM collections and scientists but also by other facilities (e.g. BRCs, microbiological laboratories) that are interested in the same topics.

Keywords

Genomics; BCCM; microorganisms; genomes; culture collection; bioinformatics; taxonomy; phylogeny; molecular evolution; biodiversity.

1. INTRODUCTION

The Belgian Coordinated Collections of Microorganisms (BCCM) is a unique initiative of the Belgian Federal Government created in 1983 by the Belgian Science Policy Office. BCCM is a Biological Resource Centre (BRC) that preserves and provides microbial and genetic resources to support life sciences and the biotechnology sector in the field of fundamental and applied research. Currently, BCCM is a consortium of seven public microbiological collections acting under the central organization of a coordination cell (www.bccm.belspo.be). The major missions of BCCM are to preserve and distribute microbial diversity, to share experience in microbial cultures and management, to give access to strain information and catalogues (through a common website), to support research on microorganisms and their function in ecosystems, and to enhance the cooperation at the international level. The consortium as a whole operates under an ISO 9001 certification since 2005.

The BCCM consortium/collections participate to different initiatives for international cooperation and networking such as the EC research infrastructure MIRRI (Microbial Resource Research Infrastructure, www.mirri.org). It is also an active member of international organisations including the European Culture Collections' Organisation (ECCO, www.eccosite.org) and the World Federation for Culture Collections (WFCC, www.wfcc.info).

Prokaryotic and eukaryotic microorganisms represent most of the diversity present on Earth and can be found in every environment capable of supporting life. They play a major role in a huge range of functions, from ecosystems to humans, and contribute to countless applications. Nevertheless, the current knowledge in microbiology barely represents the tip of the iceberg and microbiological research still needs to tackle the considerable task of finding and understanding the yet undiscovered microbial capabilities. Culture collections play an essential role in this endeavor by isolating, cultivating, identifying, preserving and distributing the cultivable diversity. In that respect, they are recognized as key hubs to ensure the FAIR use of microorganisms in research (Becker et al. 2019).

The study, valorisation and documentation of the biological resources need to stay up to date with the current developments in microbiology, the most important being whole genome sequencing that provides the entire genetic information of an organism. This is nowadays facilitated by the advent of inexpensive and high-throughput technologies that are continuously improving. These progresses are matched by advances in bioinformatics analyses of the resulting ‘big data’ in order to perform genome assembly and annotation as well as comparative analyses. Such genomic analyses can be used in both fundamental and applied research. In microbiology, genome annotation can for example reveal potential functionalities and applications of the strains. The detection in the genomes of mutations or other genetic modifications that account for specific traits like antimicrobial drug resistance is another possibility. Studies on phylogeny and taxonomy of microbial lineages are also greatly improved by the use of genomics. Taxonomy is much more than assigning a name, it is a cornerstone of biological research. During history, prokaryotic and eukaryotic taxonomy have evolved from a phenotypic approach towards a molecular framework, based on the evolutionary relationships of the organisms. Genomes are the most complete evolutionary records that can be used to infer phylogeny of organisms with high resolution, as shown by the recent proposal of a standardized bacterial taxonomy based on genome phylogeny (Parks et al. 2018).

2. STATE OF THE ART AND OBJECTIVES

Since its creation in 1983, almost 40 years ago, BCCM developed and maintained international leadership among European BRCs through the implementation of a number of features, including:

- the recognition as an international depositary authority by the World Intellectual Property Organization (WIPO) for the preservation of biological material in the frame of patent applications.
- the certification of a multi-site quality management system according to the ISO 9001 standard.
- the creation of a BCCM website and online catalogue.
- the setup of the BCCM laboratory information management system which is a common IT-platform to support, centralize and trace the daily internal operations of the different entities belonging to the BCCM consortium.

The introduction of modern technologies such as Sanger DNA sequencing or MALDI-TOF mass spectrometry was also essential to analyse microorganisms in an efficient manner. It was required to apply correct taxonomy to the microbial strains preserved at BCCM and to meet the needs of the users of the collections. Likewise, the current development and expansion of a range of technologies generally referred to as ‘next-generation sequencing’ (NGS) appeared as a new mandatory step for BRCs to answer tomorrow’s challenges in the field of microbiology. Indeed, these new tools provide high-throughput data and therefore greatly facilitate access to genetic information such as whole genome sequences or gene expression of a microbial cell under specific conditions (e.g. transcriptomics). Hence, NGS opens a new area of scientific research called genomics that was previously hindered by technological obstacles. Acquiring know-how in NGS and genomics is thus crucial for BCCM in order to remain a major BRC, for future national and international collaborations, to further characterize its strains and to answer upcoming research questions.

The first objective of the BCCM GEN-ERA project was therefore to implement expertise in genomics in the BCCM collections. This implied for some taxa the introduction of new DNA extraction protocols (DNA quality and quantity need to be higher for NGS than for Sanger sequencing), yet the real challenge was on the handling and analysis of NGS (big) data. As an example, the genome of a fungal strain contains 40 million base pairs on average and can yield more than 2 Go of data. Its analysis thus requires the installation of specific bioinformatics structures (hardware) and software for which the BCCM scientists had to be trained to ensure a long-term implementation. Importantly, of the several complementary applications offered by NGS to investigate microbial strains, whole genome sequencing was chosen primarily in the BCCM GEN-ERA project because the determination of the gene repertoire of microorganisms allows many developments, notably in molecular taxonomy, evolutionary processes and microbial metabolism. This supports the leading expertise of the collections in the field of taxonomy and phylogeny while allowing potential functional analyses. Moreover, offering strains having their whole genome sequenced is an added value for the visibility and attractiveness of the BRCs since the genome sequences of strains are increasingly requested in many disciplines.

The BCCM GEN-ERA project also aimed at answering specific research questions covering the microbial diversity of the BCCM collections (i.e., bacteria, mycobacteria, cyanobacteria, yeasts and moulds) and more specifically on selected microorganisms having a societal impact (i.e., associated with pollinating insects, causing pathologies, producing bioactive compounds or environmentally relevant). The project implicates five out of the seven BCCM collections and is a collaboration with the laboratory of Eukaryotic Phylogenomics of the University of Liège (Prof. Denis Baurain) that brings the necessary expertise in bioinformatics and (phylo)genomics.

3. METHODOLOGY

Two different bioinformatics infrastructures were envisaged for the handling and analysis of sequencing data, namely Galaxy and Nextflow. Both were tested and compared for their respective performance, appropriateness, user-friendliness and compliance with the FAIR principles (i.e., Findable, Accessible, Interoperable, Reusable).

3.1 Galaxy

The Galaxy project (https://galaxyproject.org) is a bioinformatics web platform whose objective is to make analyses, such as genomics and proteomics, available to all researchers, even with little computing skills (Afgan et al. 2018). The graphical user interface (GUI) of the web platform indeed permits researchers to have access to bioinformatics tools without the need of command lines. Although the public Galaxy servers are available to a large community of researchers, the possibility to develop studies on these servers is very limited. In particular, the number of analyses and their weight (in terms of CPU or RAM usage) on these servers represented a limitation for the research done at BCCM. The Galaxy project nevertheless offers the possibility to deploy a private Galaxy server on an HPC cluster. Since the GEN-ERA project financed a private computational node on the NIC5 HPC cluster, a cluster hosted at the University of Liege and maintained by the CECI "Consortium des Equipements de Calcul Intensif", the deployment of a "BCCM" Galaxy instance was tested.

The Galaxy project provides files configured for pre-installation of a new instance through the Ansible software (https://galaxy.ansible.com/galaxyproject). Ansible-Galaxy allows to deploy the web servers on a cluster while controlling and maintaining the system configuration in human readable files. This presents the advantage to host the full settings on a GitHub repository and to re-deploy the instance when it fails. However, the deployment on our last-generation Centos 8 HPC cluster required very specific configuration files. We notably observed security gaps in the Ansible-Galaxy files, linked to the opening of the income port (port 4632) of the HPC cluster for the web server. We therefore had to add a virtual machine (VM), hosted at the University of Liege, which complicated the installation procedure. The objective of this VM was to have the 4632 income port listening to the user's request made through the GUI and to send the analyses (and corresponding files) to the HPC cluster through Pulsar (https://github.com/galaxyproject/pulsar). Pulsar was deployed on the VM during the installation of the Galaxy server and manually connected on the HPC cluster by the system administrator from NIC5. It was used as a connection between the users and the HPC system. The installation of Pulsar required the editing of more than 100 variables of Galaxy-Ansible files and was very specific to the couple made by NIC5 and the VM. Despite the usage of Ansible-Galaxy, the tedious installation of Pulsar was only half reproducible since it was not possible to use Ansible-Galaxy on NIC5. With Pulsar and the web-servers installed, the Galaxy instance was nearly complete. The remaining step was to manually reinforce the VM security to permit the usage of a VPN with Galaxy, which is not supported by Ansible-Galaxy. In the end, due to the complexity of this installation process, it appeared that a system administrator would be needed to maintain the infrastructure.

Although not fully reproducible and time-consuming to maintain, our Galaxy servers offered interesting services to the BCCM community. The presence of the GUI was intuitive and adapted to users lacking skills in bioinformatics. Galaxy also offers the possibility to create common workflows, modifiable by the users, and databases. Although it was possible to increase the reproducibility by maintaining the tools and workflows with Ephemeris (https://github.com/galaxyproject/ephemeris), a tool able to automatically install programs on the servers based on a list, the workflows were not FAIR. Indeed, the tools were clearly not interoperable or reusable because of the difficulties linked to the Galaxy installation.

The utilization of the Galaxy servers was tested. It demonstrated that some mandatory programs required for modern bioinformatics practices (e.g. Anvi'o, RagTag, Mantis, BRAKER2, SCaFoS, Bio-Must-Core, EUKcc) were not available in the installation system of Galaxy. These tools were too recent or simply not added into the Galaxy project at their publication. On the other hand, the mosaic of available tools in Galaxy complicated the creation of the workflows. For these reasons but also due to the necessity of a system administrator to maintain the instance and the lack of FAIR practices, we decided to continue the development of the workflows with Nextflow, built in parallel with the Galaxy instance.

3.2 Nextflow

3.2.1 Introduction

Nextflow is designed to perform bioinformatics using command lines instead of a GUI. For the BCCM GEN-ERA project, 14 Nextflow workflows, sustained by 11 Singularity containers (i.e., boxes where software and dependencies are encapsulated) were implemented. They cover the common genomics needs for microbial taxonomy and metabolic modelling of the BCCM collections. They can be used to infer comparative genomics and metabolic analyses on prokaryotes and small eukaryotes, in a completely reproducible manner. Each workflow is accompanied by a python script for parsing and formatting results, included in the container. The workflows are provided to the users as a program and include a help section. They can be run with a single command line, increasing the reproducibility of the analyses. The databases used by the different workflows are automatically downloaded at the first run of the workflow if they are not provided. This GEN-ERA toolbox (workflows, Singularity definition files, companion scripts) is freely available from the GitHub repository: https://github.com/Lcornet/GENERA. This repository includes a large documentation for each tool, focusing notably on HPC usage (see point 5.1).

3.2.2 Genomes related workflows

The first four workflows are related to genome acquisition and annotation. The first tool, ***Genome-downloader.nf***, creates a mirror copy of the NCBI taxonomy (Federhen 2012, Scoch et al. 2020) and downloads the genomes according to this taxonomy. The user should specify the name of the group and the taxonomic rank (for instance, “Gloeobacterales” and “order”). The specification of the taxonomic rank makes ***Genome-downloader.nf*** resilient to future changes of the NCBI taxonomy. The second tool, ***Assembly.nf***, is dedicated to genome production. This workflow can assemble genomes and metagenomes from Illumina short reads as well as PacBio or Nanopore long reads data thanks to the use of SPAdes (Bankevick et al. 2012), metaSPAdes (Nurk et al. 2017) and metaFlye (Kolmogorov et al. 2020). An option for metagenomic binning, with MetaBAT2 (Kang et al. 2019) and CONCOCT (Alneberg et al. 2013), is provided. These two binners algorithms are complementary since CONCOCT is more efficient for eukaryotic data while MetaBAT2 is pre-trained for prokaryotic sequences. The third tool related to genomes is ***GENcontams.nf*** for genomic contamination estimation and genomes statistics production. Estimation of genomic contamination (i.e., the inclusion of foreign DNA in a genome assembly) requires the use of multiple tools to better catch contaminants (Cornet & Baurain 2022). Indeed some tools are dedicated to bacterial genomes (CheckM [Parks et al. 2015], GUNC [Orakov et al. 2021]), others are specific to eukaryotes

(EukCC [Saary et al. 2020]), while some can work on both domains without the ability of interdomain detection (BUSCO [Manni et al. 2021]). In addition, Physeter (Lupo et al. 2021) and Kraken2 (Wood et al. 2019) are two tools able to perform interdomain detection, allowing for instance the detection of eukaryotic contamination in bacteria (and vice versa). To facilitate the detection of contaminants, all these tools are implemented in ***GENcontams.nf***. The last tools of this section are related to genome annotation (i.e., protein prediction). Although bacterial protein annotation is automatic in the different GEN-ERA workflows, we provide a Singularity container for Prodigal (Hyatt et al. 2010). For eukaryotic annotation, the ***AMAW*** (Meunier et al. 2022) and ***BRAKER.nf*** workflows were developed to be used on eukaryotic genomes. ***BRAKER.nf*** is able to download RNAseq evidence, based on a user list, and to use proteins from OrthoDB (Zdobnov et al. 2021) to annotate genomes with BRAKER2 (Bruna et al. 2021). ***AMAW*** automatizes evidence collection based on the species name (Meunier et al. 2022) and is dedicated to annotation of non-model organisms.

3.2.3 Phylogeny related workflows

The section covers phylogenomics analyses from the orthology inference to the production of phylogenetic trees. The first workflow, ***Orthology.nf***, is related to orthology inference. Bacterial genomes (or proteomes) and eukaryotic proteomes are the basis of ***Orthology.nf.*** Two software can be used to compute proteins orthologous groups (OGs), OrthoMCL (Li et al. 2003), available for prokaryotes only, and OrthoFinder (Emms & Kelly 2019), available for both domains. ***Orthology.nf*** automatically provides the core genes, shared by all the organisms in unicopy, and the specific genes found only in a series of organisms (based on a user-provided list). The proteins OGs can be further enriched with orthologous from new organisms by ***OGsEnrichment.nf*** using Forty-two (Irissary et al. 2017). The proteins OGs can also be reverse translated by ***OGsRtranslate.nf*** using leel (Rodríguez et al. 2017). The proteins and nucleotides OGs can both be used for phylogenomics analyses with ***Phylogeny.nf***. This workflow computes phylogenomics inference using BMGE (Criscuolo & Gribaldo 2010) for unambiguously aligned site selection, ScaFoS (Roure et al. 2007) for sequence concatenation, and RAxML (Stamatakis et al. 2008) for phylogenetic inference. With an interface very similar to ***Phylogeny.nf***, proteins and nucleotides OGs can be provided to ***PhylogenySingle.nf*** to compute individual gene trees with RAxML (Stamatakis et al. 2008). The last tool of this section is ***ORPER.nf***, which was published independently (Cornet et al. 2021), and is designed to constrained SSU rRNA phylogeny with RAxML (Stamatakis et al. 2008). This tool first produces a phylogenomic analysis using ribosomal proteins and constrains SSU rRNA phylogeny using this reference phylogenetic tree. This multi-locus constraint is used to reduce the inaccuracy of single-gene analysis (Cornet et al 2021).

3.2.4 Other workflows including metabolic modelling

Three remaining workflows are also provided in the GEN-ERA toolbox. The first one, ***ANI.nf***, computes average nucleotide distance between genomes using fastANI (Jain et al. 2018). The second one, ***GTDB.nf***, uses GTDBtk (Chaumeil et al. 2020) to classify prokaryotic genomes according to the Genome Taxonomy Database (GTDB) taxonomy (Parks et al. 2019, Parks et al. 2021). The last workflow, ***Metabolic.nf***, is dedicated to protein function annotation using Mantis (Queiros et al. 2021) and metabolic modeling of prokaryotes using Anvi'o (Eren et al. 2015) with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kaneisha & Goto 2000).

➔ For more details: Cornet et al. 2022. The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics. Submitted to GIGAsciense.

3.3 Advantages and disadvantages of Galaxy and Nextflow

TABLE I recapitulates and compares the features and requirements of Galaxy and Nextflow, based on our tests and observations. It appeared that Nextflow fulfilled most of the criteria for a long-term and FAIR utilization of the bioinformatics infrastructure at BCCM. The only disadvantage, as compared to Galaxy, is the user-friendliness. The use of command lines is indeed less intuitive and necessitated dedicated trainings and documentation. However, the number of command lines could be reduced to a minimum thanks to the Singularity containers.

TABLE I. Pros and cons of Galaxy and Nextflow

Feature	Galaxy	Nextflow
Installation	not reproducible, system-specific	reproducible
System administrator	need after the project	need only for workflow creation
Programs availability	lack of programs	no limitation
Workflow	modifiable	fixed, with conditional usage (option)
Workflow publication	difficult	easily publishable
Training	limited, very intuitive GUI	required, use of command lines
FAIR practices	no	yes, easily exportable

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

The Nextflow infrastructure implemented at BCCM in collaboration with the University of Liège was used to investigate case studies on the different types of microorganisms preserved by the participating collections. The latter selected species and strains with a significant societal impact and pending research questions that could be addressed by genomic analyses.

4.1 BCCM/IHEM: *Trichophyton* genomes to address dermatophytes phylogeny, epidemiology and identification

Type of microorganism: filamentous fungi (moulds)

Dermatophytes are a group of highly prevalent pathogenic fungi that cause superficial skin infections in both humans and animals. Among them, species belonging to the *Trichophyton rubrum* complex are strictly anthropophilic and can infect the glabrous skin, the scalp and the nails in immunocompetent patients. They are intensively studied due to their importance in human health. Three species are currently accepted in this complex: *T. rubrum*, *T. violaceum* and *T. soudanense* (Graser et al. 2000, de Hoog et al. 2017, Packeu et al. 2020). *T. rubrum* has a worldwide distribution and is increasingly prevalent in North America, Europe, Australia and East Asia since the 1950s following a change in habits, such as the use of occlusive footwear. *T. violaceum* is predominant in Middle Eastern countries, East Africa and South China while *T. soudanense* is endemic to West African countries. These three species also vary in their morphology and the type of infection that they cause.

Genetically, they are closely related and the internal transcribed spacer (ITS) of the ribosomal region, although currently considered as the most informative marker for molecular delineation, is largely conserved between them (Graser et al. 2000, de Hoog et al. 2017, Zhan et al. 2018). Interestingly, this low genetic variation of the ITS region contrasts with a high phenotypic variability. This resulted in the description of taxa that were later synonymized with the current three species (Graser et al. 2000, de Hoog et al. 2017, Su et al. 2019):

- *T. rubrum* includes *T. raubitschekii*, *T. kanei*, *T. fischeri*, *T. flavum, T. fluviomuniense, T. pedis, T. rodhainii, T. kuryangei* and *T. megninii*
- *T. soudanense* includes *T. circonvolutum* and *T. gourvilii*
- *T. violaceum* includes *T. yaoundei, T. glabrum* and *T. violaceum var*. *indicum*

Phylogenetic analyses also revealed two distinct clusters of strains that did not group with any of the three accepted species, namely a cluster around the *kuryangei* morphotype and a cluster around the *megninii* morphotype (Packeu et al. 2020). However, further analyses were required to determine their exact position and status. Similarly, the *T. violaceum* morphotype *yaoundei*, remained to be deciphered as well.

The intraspecific variation of *T. rubrum* genomes is extremely low. A previous study showed that seven *T. rubrum* strains sampled across the globe, displayed a 99.99% identity, which is indicative of the clonal nature of this species (Persinoti et al. 2018). The close genetic distances also suggest that the species in the *T. rubrum* complex diverged in a very short evolutionary time span. Such recent divergence from a common lineage can be incomplete, resulting in species boundaries that are difficult to draw (Su et al. 2019).

Groups with distinct morphology, geographical distribution and clinical aspects are thus well-known within the *T. rubrum* complex despite the low genetic variation. Therefore, our study aimed to provide additional support for a stable taxonomy through phylogenomic analyses. Hence, we tested the confirmation of *T. rubrum*, *T. violaceum* and *T. soudanense* and investigated the possible reinstatement of *T. megninii*, *T. kuryangei* and *T. yaoundei* as separate species. This work represented the first phylogenomic analysis of the *T. rubrum* complex.

Twenty-seven strains belonging to the *T. rubrum* complex were selected for the analyses including 19 from the BCCM/IHEM fungi collection and eight strains for which the assembled genomes was downloaded from GenBank. In addition, two strains of *T. interdigitale* were selected as the outgroup. Isolates were chosen to maximize the diversity within the *T. rubrum* complex, including representatives of a wide variety of morphotypes (*megninii*, *kuryangei*, *yaoundei*, *kanei*, *fischeri* and *raubitscheckii*). After genome assembly and annotation, the phylogeny was inferred by maximum likelihood on a dataset of 3105 core genes. Two independent methods, namely bootstrapping and jackknife, were used to assess the robustness of the phylogenomic results. The latter revealed a highly resolved phylogenomic tree with six separate clades (Figure 1). *Trichophyton rubrum*, *T. violaceum* and *T. soudanense* were confirmed in their status of species. The morphotypes *T. megninii, T. kuryangei* and *T. yaoundei* all grouped in their own respective clade with high support, suggesting that these morphotypes should be reinstituted to the species-level.

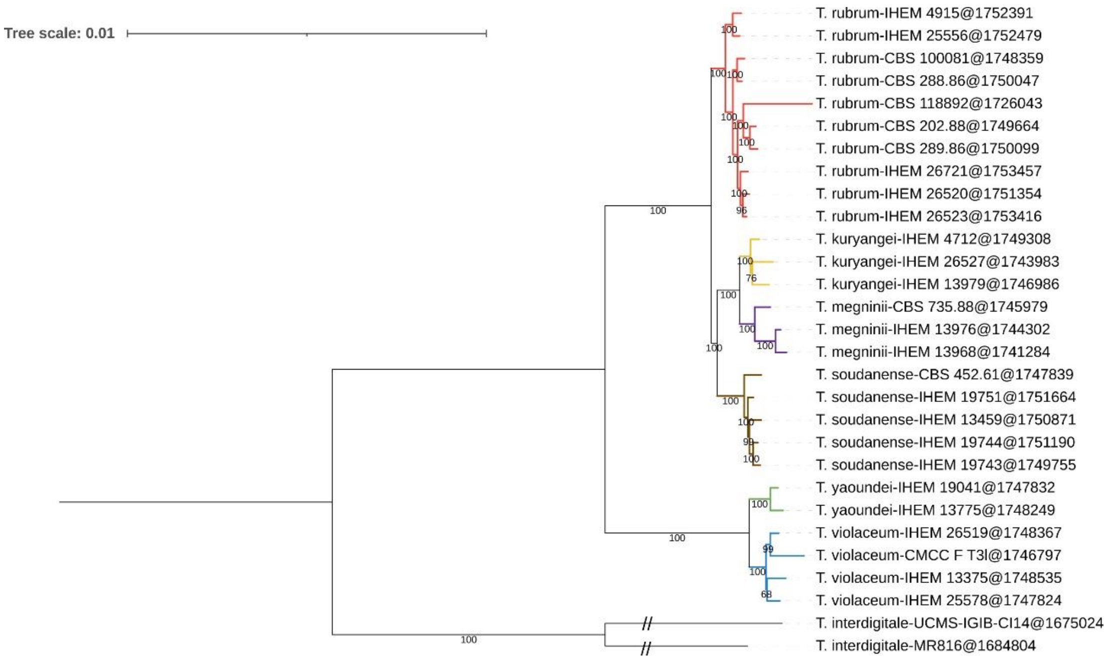

Tree scale: 0.01
T. rubrum-IHEM 4915@1752391
T. rubrum-IHEM 25556@1752479
T. rubrum-CBS 100081@1748359
T. rubrum-CBS 288.86@1750047
T. rubrum-CBS 118892@1726043
T. rubrum-CBS 202.88@1749664
T. rubrum-CBS 289.86@1750099
T. rubrum-IHEM 26721@1753457
T. rubrum-IHEM 26520@1751354
T. rubrum-IHEM 26523@1753416
T. kuryangei-IHEM 4712@1749308
T. kuryangei-IHEM 26527@1743983
T. kuryangei-IHEM 13979@1746986
T. megninii-CBS 735.88@1745979
T. megninii-IHEM 13976@1744302
T. megninii-IHEM 13968@1741284
T. soudanense-CBS 452.61@1747839
T. soudanense-IHEM 19751@1751664
T. soudanense-IHEM 13459@1750871
T. soudanense-IHEM 19744@1751190
T. soudanense-IHEM 19743@1749755
T. yaoundei-IHEM 19041@1747832
T. yaoundei-IHEM 13775@1748249
T. violaceum-IHEM 26519@1748367
T. violaceum-CMCC F T3l@1746797
T. violaceum-IHEM 13375@1748535
T. violaceum-IHEM 25578@1747824
T. interdigitale-UCMS-IGIB-CI14@1675024
T. interdigitale-MR816@1684804
100
100
100
100
100
100
100
100
96
100
76
100
100
100
100
100
100
99
100
100
100
99
100
68
100
//
//

Figure 1. Large-scale protein analysis of 27 strains belonging to the *T. rubrum* complex. The maximum likelihood tree was inferred on 3105 core genes under the PROTGAMMALGF model with RAxML from a supermatrix of 29×1.688.754 unambiguously aligned amino-acid positions. *T. rubrum* is in red, *T. kuryangei* in yellow, *T. megninii* in purple, *T. soudanense* in brown, *T. yaoundei* in green and *T. violaceum* in blue. Bootstrap support values are shown at the nodes. The branch length of the outgroup is divided by five.

In addition, we mined core genes using the Robinson-Foulds distance comparison approach in order to identify new candidate markers for future phylogenetic analyses and species identification. It showed that a combination of two markers (a ubiquitin-protein transferase and a MYB DNA-binding domain-containing protein) can mirror the phylogeny obtained using genomic data, and thus represent potential new markers to accurately distinguish the species belonging to the *T. rubrum* complex. These new gene markers would resolve the *T. rubrum* complex phylogeny better than conventional barcodes (i.e. ITS), which would facilitate future identifications of isolates, and (epidemiological) studies on this complex.

➔ For more details: Cornet et al. 2021. The taxonomy of the *Trichophyton rubrum* complex: a phylogenomic approach. Microbial Genomics 7: 000707.

4.2 BCCM/ITM: *Mycobacterium* genomes to address their virulence and interactions with host immune cells

Type of microorganisms: mycobacteria

The prevalence of diseases caused by non-tuberculosis mycobacteria (NTM) is increasing worldwide and their diagnosis is still very challenging for clinicians. Indeed, due to the continuous exposure of individuals to environmental NTM, it is difficult to determine whether an NTM isolated from a clinical specimen represents a colonization or is a true etiology agent of the disease. The first step in clinical diagnosis requires the correct identification of the isolated NTM species. For research purposes, it is also important to better understand the phenotypes and pathogenicity characteristics of various NTM species. Good knowledge of their genomes is therefore paramount.

More than 50 different NTM species are now considered to be opportunistic pathogens in humans, and the number is rising. Despite the growing demand for genomic information for identifying NTM at the subspecies level, there is still a lack of information in terms of sequences and specific features of these species in public databases. By assembling and annotating newly sequenced NTM genomes, the goal was to substantially fill these gaps and increase awareness of these NTM species.

This work aimed to provide good quality genomes to the public database of 16 NTM species (TABLE II) and assessing the genomic features of these species. These genomes were newly determined and fall into different branches of the phylogenetic tree constructed by Tortoli et al. (2017).

TABLE II. The selected NTM species for genome sequencing, assembly and annotation and their potential clinical relevance.

NTM species	Potential clinical relevance
M. iranicum	Chronic pulmonary infection, cutaneous lesions
M. triviale	Pulmonary infection
M. nonchromogenicum	Pulmonary infection
M. paraense	Pulmonary infection
M. persicum	Pulmonary infection
M. alsense	Pulmonary infection
M. fragae	Pulmonary infection
M. szulgai	Skin infection, Pulmonary infection
M. heraklionense	Tenosynovitis
M. angelicum	Type II respiratory failure
M. engbaekii	No case
M. longobardum	Osteomyelitis
M. palustre	Pediatric lymphadenitis
M. europaeum	Cavitary pneumonia, cervical lymphadenitis
M. riyadhense	Pulmonary infection
M. diernhoferi	No case

Draft assembly were generated by long reads obtained from the PacBio (Sequal II) sequencing. The Sequal II uses the “single molecule real-time” (SMRT) technology designed to boost the contiguity, completeness and correctness of a draft assembly, as long reads can better cover high complex regions such as high repetitive regions as compared to Illumina short reads. SMRT has evolved to a different type of long read, known as highly accurate long reads, or HiFi reads, resulting in a higher accuracy comparable with short reads. Therefore, there is also no bias introduced via DNA amplification stages. But even with the similar error rate between the Pacbio Hifi reads and Illumina short reads, they remain different sequence technologies that suffer from different types of errors. Therefore, we opted in a later stage to perform additional Illumina short read sequencing to polish the draft long read assembly and improve accuracy.

For assembly, it was originally planned to use the innate PacBio end to end SMRTlink software, since it supports both PacBio Hifi reads and allows for hybrid assemblies using short and long reads. Nevertheless, these options were not supported by the *Assembly.nf* workflow of GEN-ERA. Moreover, the SMRTlink software could not be used since it is proven to be very computer resource intensive, which was beyond the available server capabilities. We therefore opted for a local installation of the individual PacBio tools distributed via Bioconda, intended for command line users to build a pipeline that mimics the SMRTlink software in a conda environment (https://github.com/PacificBiosciences/pbbioconda). In short, after demultiplexing, raw Bam files were used to generate HiFi reads in fastq format. Then the CCS tool from the pbbioconda was used to combine multiple subreads of the same SMRTbell molecule. Different draft assemblies were generated with Flye (Kolmogorov et al. 2019) and IPA (the official PacBio software for HiFi genome assembly), and raw subreads were polished with the pbmm2 tool to perform some read correction. Some final adjustments were applied by examining the circularity and the strand orientation of the assembly and to set the origin of replication. As the NTM species were expected to have circular chromosomes, the circularity of genome was checked by looking at overlap between the beginning and end of genome sequences using the Circlator tool (Hunt et al. 2015).

For the quality controls, various assembly metrics reflecting various aspects of the assembly were used to estimate the accuracy and the quality of the assemblies (TABLE III). The assembled genomic DNA sequences were annotated using the Prokka pipeline v1.14.6 (Seemann 2014) which uses external feature prediction tools (such as Prodigal, RNAmmer, Aragorn, SignalP, and Infernal) to help identify genomic features within the contigs. A BLAST of the extracted 16S genes was performed to do a consensus comparison with the NCBI database.

As seen in TABLE III, decent de novo assemblies were obtained for 12 out of 16 newly sequenced NTM strains. To further improve the accuracy of the assemblies, an additional polishing step will be included with Illumina short reads, which were generated only very recently. In addition, the quality of the genomes will be examined through the *GENcontams.nf* workflow of the GEN-ERA toolbox. The final genomes will then be analyzed for the presence of special genomic features linked to their pathogenicity such as plasmids, phage, gene gain and lost evolutionary profile.

TABLE III. Parameters to evaluate quality of assembly achieved. Red values indicate failure to pass assembly QC measurements.

NTM species (strain number)	Number of contigs	Circular	Largest polished contig (Mb)	Complete BUSCO (%)	NCBI 16S rDNA BLAST species
M. alsense (ITM 500938)	1	yes	5,74	99,5	*Mycobacterium alsense* strain TB 1906T
M. europaeum (ITM 500936)	2	yes	5,64	99,2	*Mycobacterium europaeum* strain DSM 45397
M. heraklionense (ITM 500930)	1	yes	5,06	99,4	*Mycobacteriuzm heraklionense* strain JCM 30995
M. szulgai (ITM 500083)	2	yes	7,02	99,4	*Mycobacterium szulgai* strain DSM 44166
M. triviale (ITM 500028)	1	yes	3,64	99,3	*Mycolicibacillus trivialis* strain ATCC 23292
M. angelicum (ITM 500927)	1	yes	6,73	99,2	*Mycobacterium szulgai* strain JCM 18264
M. paraense (ITM 500925)	1	yes	5,63	99,2	*Mycobacterium paraense* strain IEC23
M. iranicum (ITM 500933)	3	yes	6,33	99,5	*Mycobacterium iranicum* strain H39
M. fragae (ITM 500939)	1	yes	4,77	99,2	*Mycobacterium fragae* strain HF8705
M. nonchromogenicum (ITM 500034)	9	yes	6,17	99,8	*Mycolicibacterium parafortuitum* strain JCM 6367
M. engbaekii (ITM 500921)	1	yes	4,58	99,4	*Mycolicibacter engbaekii* strain ATCC 27353
M. diernhoferi (ITM 500012)	1	yes	6	99,1	*Mycobacterium diernhoferi* strain ATCC
M. longobardum (ITM 500934)	4	no	4,36	93,8	*Mycolicibacter longobardum* strain DSM 4539
M. palustre (ITM 500931)	1	no	4,22	31,5	*Mycobacterium palustre* strain E846
M. persicum (ITM 500926)	3	no	6,04	94,7	*Mycobacterium persicum* strain AFPC 000227
M. riyadhense (ITM 500928)	63	no	0,04	-	Failed draft assembly

4.3 BCCM/LMG: Genomes of bumblebee gut bacteria to gain knowledge on microbiont-host interactions

Type of microorganisms: bacteria

Honeybees (*Apis* spp.) and bumblebees (*Bombus* spp.) harbor a gut microbiome that is important in health and metabolism. This microbiome is highly conserved, with 95% of the gut microbionts falling within a few phylotypes that include Actinobacteria (*Bifidobacterium*, *Bombiscardovia*), Bacteroidetes (*Apibacter*), Firmicutes (*Lactobacillus* [the so-called Firm-5 or Lacto-1 taxon], *Bombilactobacillus* [Firm-4, Lacto-2], and *Apilactobacillus* [Lacto-3]), Alphaproteobacteria (*Bartonella*, *Bombella*, *Commensalibacter*), Betaproteobacteria (*Snodgrassella*), and Gammaproteobacteria (*Frischella*, *Gilliamella*) (Martinson et al. 2011). The genera *Bifidobacterium*, *Lactobacillus*, *Bombilactobacillus*, *Gilliamella* and *Snodgrassella* are generally considered as the core microbionts of honeybees and bumblebees (Zhang et al. 2021). Knowing how this microbes interact to affect the host has been identified as a key challenge to understand bee health.

Recently, Praet et al. (2018) were able to isolate the main groups of bumblebee microbionts detected through metagenomics analyses. Our study aimed to generate genome sequences for the major bumblebee endosymbionts for which cultivated representatives are available, in order to characterize their functional potential. Functional analyses of the genomes of bumblebee gut bacteria can provide novel insights into the host-microbiont interactions and the role of these bacteria in the bumblebee gut. In addition, most bumblebee isolates obtained by Praet et al. (2018) were identified through partial 16S rRNA gene sequences and require further analyses to classify them properly. Therefore, the goal of our study was also to use genome sequence analyses to verify the identity of the selected bumblebee isolates and formally name novel species, if present.

Ninety-eight strains representing the main groups of bumblebee microbionts detected through metagenomics analyses and type strains of closely related species, were selected for initial whole genome sequence data analysis. The bumblebee isolates were chosen to maximize the diversity within each group. For 28 strains, the assembled genomes were downloaded from GenBank while for 70 strains a draft genome sequence was determined. After genome assembly, average nucleotide identity (ANI) values were calculated using the OrthoANIu algorithm to assess the identity of the bumblebee isolates. This revealed 4 novel species, *Fructobacillus* sp. nov. (7 isolates), *Lactobacillus* sp. nov. (1 isolate), *Snodgrassella* sp. nov. 1 (7 isolates) and *Snodgrassella* sp. nov. 2 (2 isolates).

Further analyses during this project focused on *Snodgrassella*. The other groups of bumblebee microbionts will be further investigated outside this project. *Snodgrassella* is a genus of Betaproteobacteria that lives in the gut of honeybees (*Apis* spp.) and bumblebees (*Bombus* spp). Since its description in 2013, a single species (i.e., *Snodgrassella alvi*), was named (Kwong & Moran 2013). The formal description and naming of this species were based on three strains: one from a honeybee gut sample (wkB2^T from *Apis mellifera*) and two from bumblebee gut samples (wkB12 from *Bombus bimaculatus* and wkB29 from *Bombus vagans*). Many additional strains have been reported since, and some phylogenetic analyses revealed a clear separation between *Snodgrassella* isolates of *Apis* spp. and those of *Bombus* spp. (Koch et al. 2013, Steele & Moran 2021), whereas others reported that the two groups were not monophyletic (Engel et al. 2014, Kwong et al. 2017).

During this project, we analyzed whole-genome sequences of 75 *Snodgrassella* strains from 4 species of honeybees and 14 species of bumblebees. For 66 strains, assembled genomes were downloaded from GenBank. Average nucleotide identity (ANI) analyses revealed seven species among the analyzed *Snodgrassella* strains, including *S. alvi*. Phylogenomic analyses based on a dataset of 254 highly conserved *Neisseriaceae* core genes showed that the strains formed a monophyletic lineage within the *Neisseriaceae* family. It also revealed monophyly of *Snodgrassella* strains isolated from bumblebees and paraphyly of strains isolated from honeybees, with strains from Asian honeybees (Figure 2, *Apis* group 1) being an early diverging group. We formally named two new *Snodgrassella* species that were isolated from bumblebees: i.e., *Snodgrassella gandavensis* sp. nov. (Figure 2, *Bombus* group 4) and *Snodgrassella communis* sp. nov. (Figure 2, *Bombus* group 5).

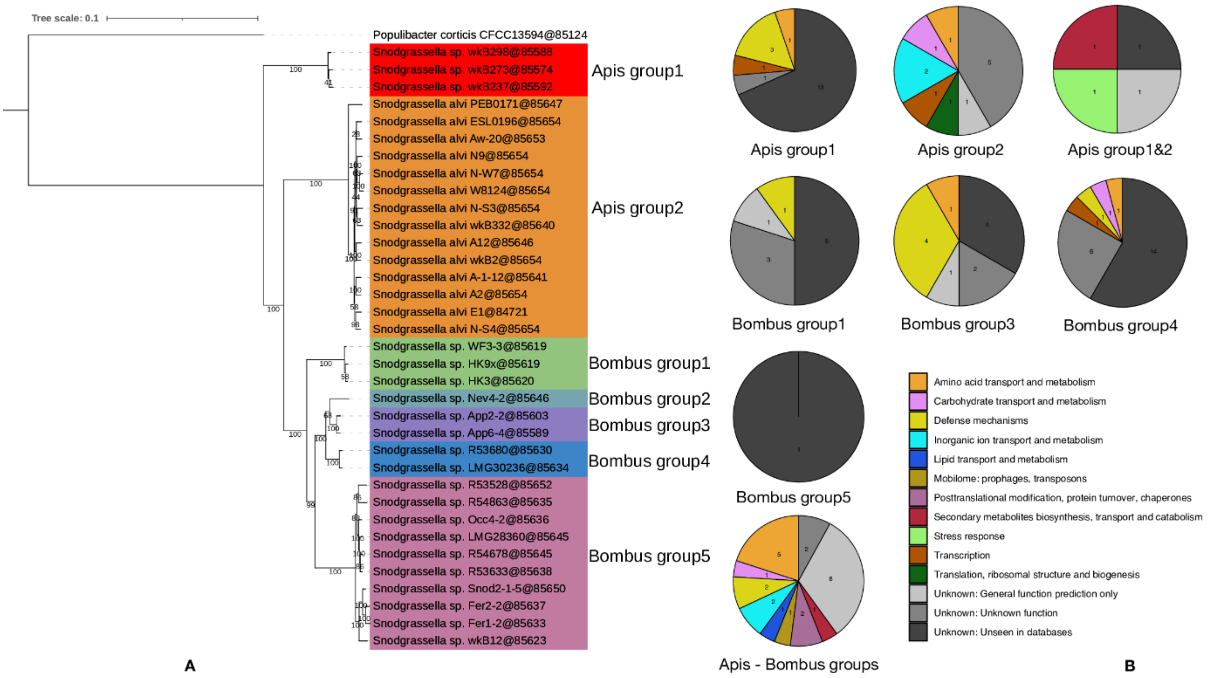


Figure 2. *Snodgrassella* phylogeny, after dereplication of highly similar genomes, and metabolic analysis per species group. (A) Maximum likelihood tree inferred on 254 core genes under the PROTGAMMALGF model with RAxML from a supermatrix of 36 organisms by 86,654 unambiguously aligned amino acid positions. Bootstrap support values are shown at the nodes. (B) Functional analyses were performed using COG and Mantis. Numbers indicated in the pie charts correspond to absolute numbers of OGs identified in the respective *Snodgrassella* subgroups. Specific genes were computed for entire groups before dereplication.

We further detected 107 genes specific to the seven *Snodgrassella* species and found no evidence for horizontal gene transfer for the large majority of these genes through examination of evolutionary scenarios. Functional analyses among these genes revealed the importance of small proteins, defense mechanisms, amino acid transport and metabolism, inorganic ion transport and metabolism and carbohydrate transport and metabolism.

- ➔ For more details: Cornet et al. 2022. Phylogenomic analyses of *Snodgrassella* isolates from honeybees and bumblebees reveals taxonomic and functional diversity. mSystems 7(3):e0150021.

4.4 BCCM/MUCL: *Starmerella* genomes to address molecular evolution, secondary metabolite production, and horizontal gene transfers

Type of microorganism: yeasts

A yeast was the first eukaryote to have its genome determined. The most intensely sequenced yeast species is the bread, beer, and wine yeast *Saccharomyces cerevisiae* (e.g. Peter et al. 2018). A considerable species diversity was covered by the analysis of 332 genomes of ascomycetous yeasts (Shen et al. 2018). This provided a robust and time-scaled phylogeny, allowed the study of numerous metabolic traits and evidenced reductive evolution as a major mode of diversification. An even larger genome phylogeny of >1000 known species of budding yeasts is in progress by the same project consortium (https://y1000plus.wei.wisc.edu/).

Many existing genome studies focus on either already industrially used species, namely *S. cerevisiae*, or on type strains. Research interest is to be extended to non-conventional yeasts with marked differences to the model yeast *S. cerevisiae*. Several members of the selected genus *Starmerella* produce sophorolipids, known as biosurfactants which are alternatives to chemical surfactants (Qazi et al. 2022). Mostly the type strain of *Starmerella bombicola* is in industrial use, its sophorolipid biosynthetic gene cluster has been investigated (Van Bogaert et al. 2013), and there are opportunities to understand further its pathway by the study of additional wild-type strains (Claus & Van Bogaert 2017, Qazi et al. 2022).

Yeasts of the genus *Starmerella* can be isolated from various species of bees and were found to dominate pollen freshly stored by honey bees (Detry et al. 2020). A beneficial relationship of bees with specific groups of bacteria is well-established (see point 4.3). While potential mechanisms of adaptation of yeasts to pollen and bees are yet unknown, yeasts have been suspected to play roles in bee nutrition or the transformation of pollen provisions. The finding that the phylogenetic clade that includes *Starmerella* and its sister genus *Wickerhamiella* (referred to as the W/S clade) carries genes of bacterial origin is of particular interest in this context (Kominek et al. 2019, Goncalves et al. 2020). Both genera appear phylogenetically separated by yeast barcode sequences (D1/D2 rDNA), although genome analyses indicate an intermediary position of at least two *Wickerhamiella* species. It appears opportune to enrich these analyses by additional taxa.

The genus *Starmerella* is of particular interest because of 1) its production of sophorolids as secondary metabolites with industrial potential, 2) a lack in understanding a potential role of yeasts in bees, 3) the presence of large numbers of genes of bacterial origin in its genomes, and 4) questions on the genetic diversity on species level.

Genomic data were obtained by the Illumina technology on 21 strains of 6 *Starmerella* species (4 strains of *S. apicola*, 3 strains of *S. apis*, 4 strains of *S. bombi*, 2 strains of *S. bombicola*, 5 strains of *S. magnolia*, 2 strains of *S. neotropicalis*, and one strain of a potential new species closely related to *S. apicola*). Fifteen of these strains are uniquely preserved at BCCM/MUCL. One strain per species for which no genome was present in public databases (*S. apis*, *S. bombi*, *S. neotropicalis*, *Starmerella* sp.) were also sequenced by the PacBio technology. Our contribution doubled the known *Starmerella* genomes.

Saccharomycetales genomes were downloaded from NCBI and a selection of genomes for the outgroup based on Goncalvez et al. 2020 was made. The sequenced genomes were assembled as axenic strains using *Assembly.nf*. Three strains sequenced by Pacbio are the best *Starmerella* genomes to date, with high contiguity and no undetermined nucleotides in the contigs. Genome quality and genomic contamination were estimated by EUkCC, Busco, and Kraken. The detected bacterial contamination might be due to mitochondria (hits on Proteobacteria), even if it is bigger than expected for fungal genomes. However, knowing that large numbers of genes of bacterial origin are present in genomes of the W/S clade, (Goncalves et al. 2018, Shen et al. 2018), it is possible that we may have detected some genomes with particularly high horizontal gene transfer frequency. Further analyses should address characteristics such as potential functions and origin of the detected bacterial contaminations, considering bacterial bee symbionts as possible source.

It was noted that the completeness levels of all *Starmerella* genomes by EUkCC, including the ones assembled with high coverage PacBio sequencing (always around 100 X) and the ones from NCBI was only around 90% (never above 92%) for *Starmerella* while it was around 98% for *Wickerhamiella* and more distantly related genomes. The values are even lower for BUSCO completeness. It is striking to see such an effect on a whole group and this may be an indication of a reduction of the genomes in this group, possibly linked to their ecology, which would be very interesting. This low completeness of the genomes needs to be further examined to test whether it is an analytical artefact or a real biological distinction. In fact, the reductive evolution in budding yeasts was suggested to be stronger in specialists like the W/S clade compared to generalist lineages (Shen et al. 2018). Specifically, after loss of alcoholic fermentation in the W/S clade postulated by Goncalves et al. 2018), the function was reestablished by acquisition of a bacterial gene in the ancestor of the *Starmerella* clade. In the distantly related budding yeast genus *Hanseniaspora* (*Saccharomycodaceae*), extensive gene losses were observed not only of typically concerned metabolic genes but also atypically of cell-cycle and DNA-damage-response genes and this in differential quantities in two lineages of the genus (Steenwyk et al. 2019).

Protein prediction of both the generated and the NCBI genomes was done as many of the NCBI genomes did not have any proteins associated. After genome annotation, the orthology tool defined 124 core genes, shared by all analysed genomes in unicopy. These were used for protein and DNA phylogenomic analyses. The DNA analyses are made in jackknife and in bootstrap, on the two first codons only and with a separate partition on the third codon, resulting in 4 DNA trees. These trees and the protein bootstrap tree (not enough core genes for protein jackknife) show a highly supported monophyletic *Starmerella* clade (Figure 3). The lowest jackknife value is 84 for a deep node of the group in a jackknife DNA tree. To make this result stronger, the phylogenetic complexity of the matrices will be estimated with a new software (Pythia), that will also be integrated into the GEN-ERA toolbox.

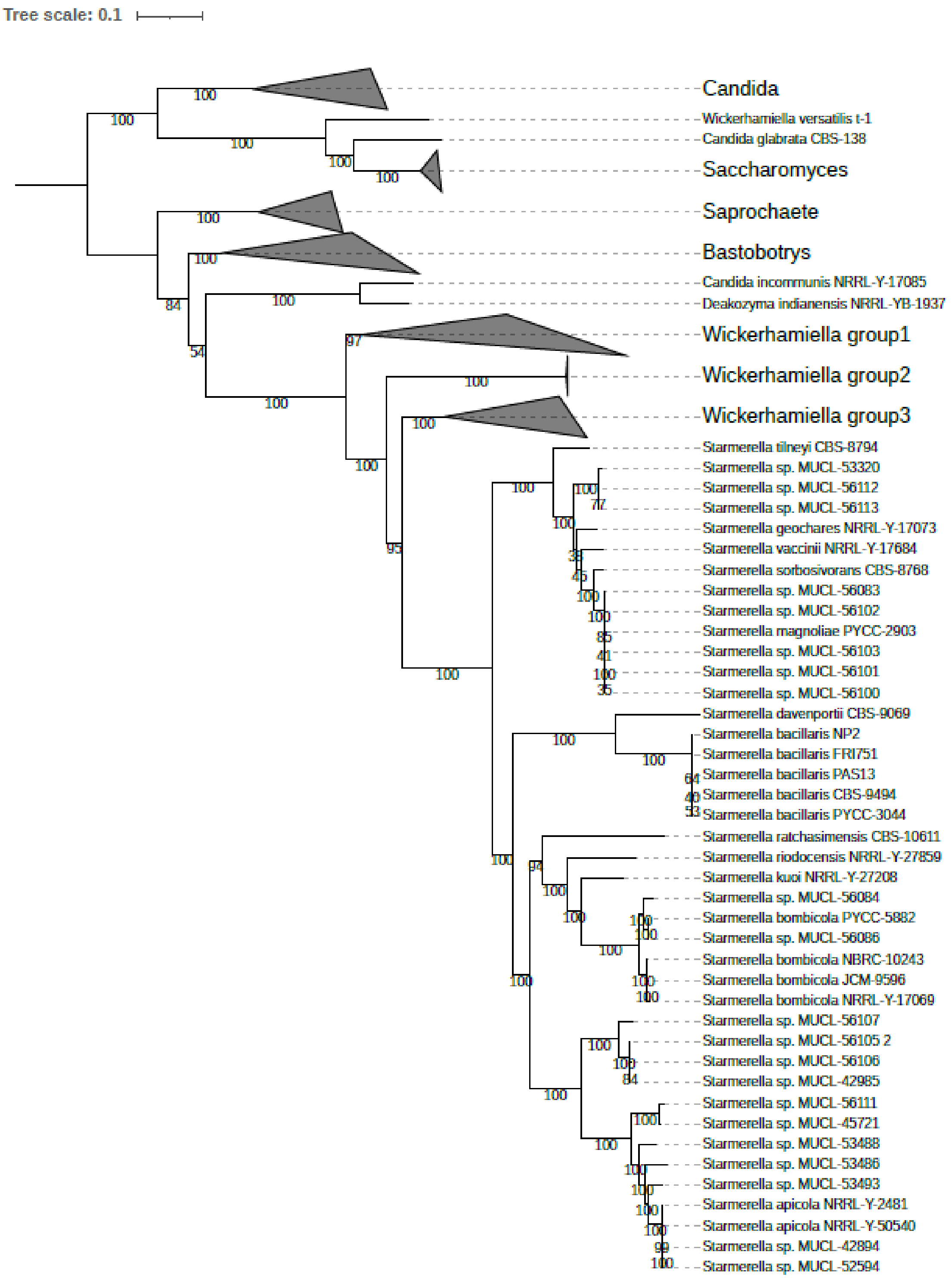

Tree scale: 0.1
Candida
Wickerhamiella versatilis t-1
Candida glabrata CBS-138
Saccharomyces
Saprochaete
Bastobotrys
Candida incommunis NRRL-Y-17085
Deakozyma indianensis NRRL-YB-1937
Wickerhamiella group1
Wickerhamiella group2
Wickerhamiella group3
Starmerella tilneyi CBS-8794
Starmerella sp. MUCL-53320
Starmerella sp. MUCL-56112
Starmerella sp. MUCL-56113
Starmerella geochares NRRL-Y-17073
Starmerella vaccinii NRRL-Y-17684
Starmerella sorbosivorans CBS-8768
Starmerella sp. MUCL-56083
Starmerella sp. MUCL-56102
Starmerella magnoliae PYCC-2903
Starmerella sp. MUCL-56103
Starmerella sp. MUCL-56101
Starmerella sp. MUCL-56100
Starmerella davenportii CBS-9069
Starmerella bacillaris NP2
Starmerella bacillaris FRI751
Starmerella bacillaris PAS13
Starmerella bacillaris CBS-9494
Starmerella bacillaris PYCC-3044
Starmerella ratchasimensis CBS-10611
Starmerella riodocensis NRRL-Y-27859
Starmerella kuoi NRRL-Y-27208
Starmerella sp. MUCL-56084
Starmerella bombicola PYCC-5882
Starmerella sp. MUCL-56086
Starmerella bombicola NBRC-10243
Starmerella bombicola JCM-9596
Starmerella bombicola NRRL-Y-17069
Starmerella sp. MUCL-56107
Starmerella sp. MUCL-56105 2
Starmerella sp. MUCL-56106
Starmerella sp. MUCL-42985
Starmerella sp. MUCL-56111
Starmerella sp. MUCL-45721
Starmerella sp. MUCL-53488
Starmerella sp. MUCL-53486
Starmerella sp. MUCL-53493
Starmerella apicola NRRL-Y-2481
Starmerella apicola NRRL-Y-50540
Starmerella sp. MUCL-42894
Starmerella sp. MUCL-52594
100
100
100
100
100
100
100
100
84
97
54
100
100
100
100
100
100
77
100
95
38
45
100
100
85
41
100
35
100
100
100
64
40
53
100
94
100
100
100
100
100
100
100
100
100
100
84
100
100
100
100
100
100
100
100
99
100

Figure 3. Large-scale protein analysis of *Wickerhamiella*/*Starmerella* clade. The maximum likelihood tree was inferred from proteins of 124 core genes. Bootstrap support values are shown at the nodes.

The phylogenetic trees derived from these data represent the largest genome-based analyses of the W/S clade. Three species were added, each with multiple strains and strains are added to another three species compared to published trees (Kominek et al. 2019, Goncalves et al 2020, Goncalves et al. 2022). This allowed to assess intraspecies variation by overall genome relatedness indices (e.g. ANI). The tree topology generally confirmed groupings known from ribosomal barcode- and genome- derived phylogenies with minor differences in *Starmerella* (*S. ratchasimensis* more closely related to a group comprising *S. bombicola* in the genome analysis than to a group of *S. bacillaris* and *S. davenportii* seen in rDNA phylogenies) and *Wickerhamiella* (W. vanderwaltii sister to W. occidentalis in the current analysis instead of *W. alocasiicola* as reported by Goncalves et al. 2022). The genome trees thereby validated the barcode sequence (D1/D2 LSU) derived phylogenies for these two genera with the limitation that the genome trees still represent a small number of species known in these genera (15 out of 48 known *Starmerella* species, 19 out of 45 known *Wickerhamiella* species).

The new genomes will be scrutinized for signs of horizontal gene transfers, known to be exceptionally frequent in species of the W/S clade (Shen et al. 2018). Identifying transferred genetic elements shall further the understanding of the molecular evolution of this group of yeasts.

Finally, the determination of additional genomes of the high sophorolipid producing species *S. bombicola* (2 new, 4 publicly available, 3 of them from the type strain) and *S. apicola* (4 new, 2 publicly available) should allow an assessment of variability in the genetic background for the production of this secondary metabolite of industrial interest (Van Bogaert et al. 2013, Qazi et al. 2022).

4.5 BCCM/ULC: Genomic selection of cyanobacteria strains with ORPER for comparative genomics of closely related Antarctica and non- Antarctica strains.

Type of microorganism: cyanobacteria

Cyanobacteria form a diversified group of prokaryotes, also called blue-green algae. This phylum is of great scientific importance, notably in terms of ecology and evolution. Cyanobacteria have indeed colonized a vast diversity of environments and are one of the major components of the phytoplankton (Rippka et al. 1979, Whitton & Potts 2012). It is within this group that oxygenic photosynthesis appeared 2.4 billion years ago, substantially raising the level of atmospheric oxygen on earth (Knoll 2003, Kopp et al. 2005, Ochoa et al. 2014). Photosynthesis has then spread into eukaryotic lineages through the primary endosymbiosis of a cyanobacterium which has led to the emergence of plasmids (Archibald 2009). Consequently, cyanobacteria is an intensively studied group resulting in a high number of genomes released in public repositories, with more than 1.000 genomes per year on NCBI during the last three year.

Due to this high sequencing rate, it is more and more difficult for culture collections to select strains for sequencing without duplicating the sequencing effort of the scientific community. To overcome this situation, we developed the Nextflow workflow ORPER, for "ORganism PlacER", which permits to determine the phylogenetic position of organisms in the genomic landscape (Cornet et al. 2021). The gold standard for the evaluation of bacterial diversity remains the SSU rRNA gene of the small subunit of the ribosomal RNA (Yarza et al. 2014), which can be easily sequenced by culture collections and is also massively available in public resources such as SILVA (Quast et al. 2013) or NCBI (Nasko et al. 2018). Nevertheless, it has been proven that single gene phylogenies, such as the one based on SSU rRNA, suffer from a bad phylogenetic resolution (Gontcharov et al. 2004, Dessimoz & Gils 2010, Lunter et al. 2008). ORPER begins by automatically downloading relevant genomes, based on a user-provided taxonomy, from GenBank (Sayers et al. 2022) or RefSeq (Nasko et al. 2018). In addition, a quality filtration of the genomes, based on genomic contamination level, is performed with CheckM (Parks et al. 2015). ORPER then infers a multi-locus reference phylogenomic tree based on ribosomal protein from the RiboDB database (Jauffrit et al. 2016). A second, and final, phylogenetic tree is then built from SSU rRNA genes only, after their extraction from public genomes. The backbone of this final tree is constrained by the multi-locus reference tree, which allows minimizing the artifact linked to single gene phylogeny (Cornet et al. 2021). The user should provide SSU rRNA sequences that will be integrated during the inference of the final tree. A dereplication of genomes and SSU rRNA is optional in ORPER. ORPER can thus select strains that are distant from any public genomes in order to increase the sequenced coverage of a taxanomic group. On the contrary, it can select strains with close representatives for comparative analyses.

We used ORPER with 152 SSU rRNA sequences from the BCCM/ULC collection, which has been deposit on NCBI servers to this end. After dereplication, the final tree produced by ORPER showec that the 140 remaining strains covered the whole cyanobacterial sequenced diversity (Figure 4), at the exception of clade 7 of Moore et al. (2019). Three BCCM/ULC strains (i.e., ULC415, ULC417, ULC381) formed a basal and not sequenced clade, which is thus of high interest for plastid emergence (Figure 4).

Thanks to ORPER and in the context of the GEN-ERA project, three strains have been selected for whole genome sequencing (ULC096, ULC102, ULC722). These strains belong to the Oscillatoriales order and are close to the genome of *Oscillatoria acuminata* (GCF_000317105.1). Two out of these three strains (ULC096, ULC102) were collected in Antarctica and the other one (ULC702) was collected in Brazil. These three strains will permit a comparative analysis and metabolic modeling of this subgroup to understand the adaptation of cyanobacteria to the polar environment. The three strains were sequenced in Nanopore and Illumina, the assembly has been done with the workflow Assembly.nf and the quality assessment with the workflow GENcontams.nf, both developed during the GEN-ERA project. It resulted in three cycled high-quality genomes, summarized in TABLE IV. The taxonomy of the genomes was determined using the Genome Taxonomy Database (GTDB) (Parks et al. 2020), with the workflow GTDB.nf. While the GTDB taxonomy indicates a species level for the

Antarctica strains, the taxonomy of ULC722 can only be determined at the family level, which indicates a rare genome in the cyanobacteria diversity.

TABLE IV: Quality assessment of genomic bins produced for BCCM/ULC strains. Assembly was done with Assembly.nf in metagenomic mode, bins were identified with GTDB.nf and quality estimated with GENcontam.nf.

Genome	Taxonomy GTDB	Number of contigs	Completeness	Contamination
ULC096	*Phormidium pseudopriestleyi*	1	100	0.32
ULC102	*Phormidium pseudopriestleyi*	1	99.99	0.68
ULC722	*Oscillatoriacea* sp.	2	99.54	0.54

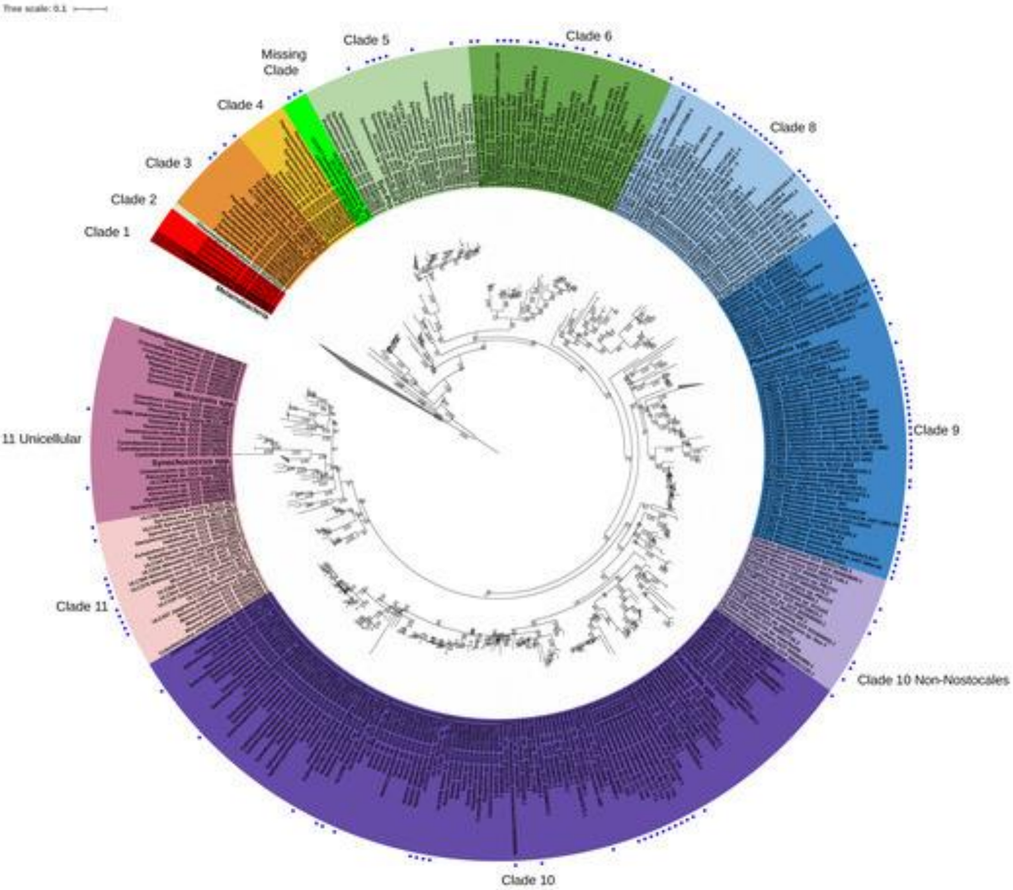


Figure 4: Constrained phylogenetic tree of cyanobacteria from the BCCM/ULC collection. The tree is the output of ORPER, a maximum-likelihood constrained inference computed under the GTRGAMMA model. Clades correspond to the groups defined in Moore et al. (2019). Clades 10 and 11 were subsequently divided into two sub-clades, resulting in the addition of respectively “Non-Nostocales” and “Unicellular” sub-clades to Moore’s phylogeny. Blue dots indicate ULC/BCCM strains. The clade absent from Moore’s phylogeny is indicated as “Missing Clade”.

➔ For more details: Cornet et al 2021. ORPER: A workflow for constrained SSU rRNA phylogenies. Genes 12: 1741.

4.6 Recommendations

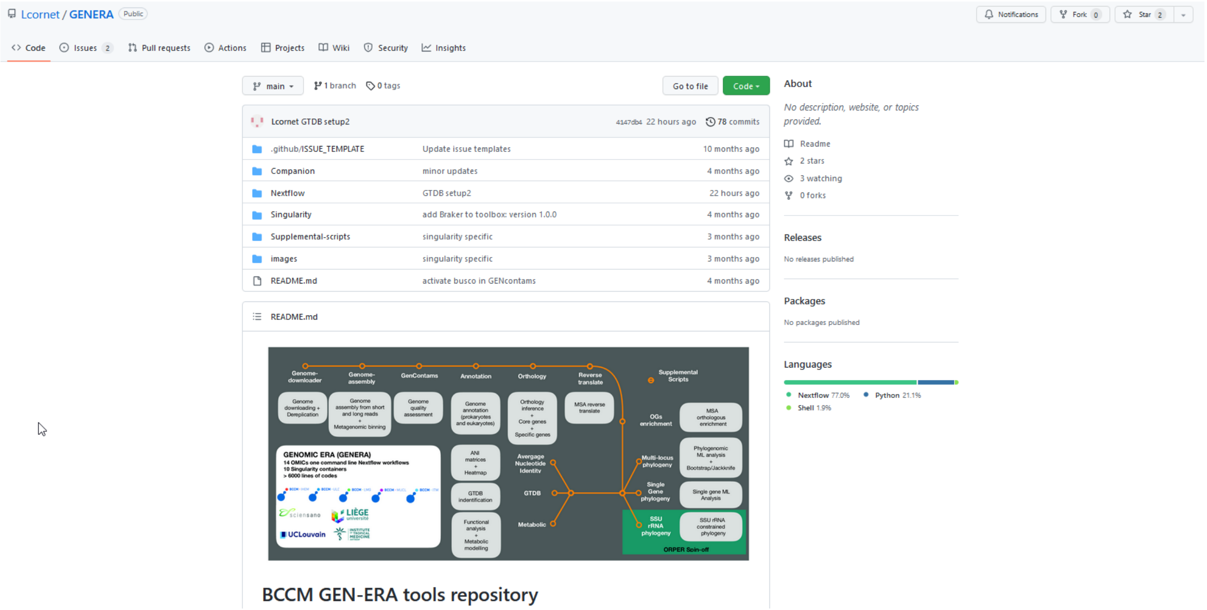
BCCM is a federal program financed by the Belgian Science Policy since 1983. Since its start, the implementation, within the collections, of necessary infrastructures (e.g. quality management system, website, online catalogue, LIMS) was supported by Belspo to enable BCCM to meet the international standards of BRCs. After 40 years of existence, BCCM now faces a new, yet essential, technological revolution, namely genomics. Genomic data allows for new areas of scientific researches in microbiology and their access is important for both BCCM and the users of the collections. It allows BCCM to remain on the map of the BRCs by increasing its international visibility and recognition. The BCCM GEN-ERA project laid the foundation of this challenge by establishing a bioinformatics framework, by bringing genomic tools and by developing genomics expertise of the BCCM scientists. This should however be seen as a starting point and a necessary investment for future analyses. Genomics should indeed become a continuous activity with the BCCM consortium. We therefore recommend the competent authorities to take this into account when evaluating the needs and missions of BCCM.

5. DISSEMINATION AND VALORISATION

The valorisation of the BCCM GEN-ERA outputs was performed following the FAIR principles. This means that all deliverables were made publicly available and can be found, accessed and reused by external users. Interoperability was also targeted by the setup of our infrastructure (e.g. Python scripts, software) in Singularity containers that can be executed on all recent operating system. The results were disseminated using various channels including publications in international open access peer-reviewed journals (see point 6), communications during conferences and a dedicated Linkedin website.

5.1 The BCCM GEN-ERA Github: an open gateway to use, reuse and learn to use bioinformatic tools for genomic analyses

As described in the methodology (see point 3), the BCCM GEN-ERA Github (https://github.com/Lcornet/GENERA) was designed to offer online and free access to analysis programs for the handling of genomic data from prokaryotic and eukaryotic microorganisms. It is a web portal that centralizes software for notably genome assembly, genome annotation, orthology inference, phylogenomics, metabolic modelling of genomes, etc. It works as a common platform that can be used by all BCCM collections and scientists but also by other facilities (e.g. BRCs, microbiological laboratories) that are interested in the same topics.



Lcornet GTDB setup2 4147db4 22 hours ago 78 commits

.github/ISSUE_TEMPLATE	Update issue templates	10 months ago
Companion	minor updates	4 months ago
Nextflow	GTDB setup2	22 hours ago
Singularity	add Braker to toolbox: version 1.0.0	4 months ago
Supplemental-scripts	singularity specific	3 months ago
images	singularity specific	3 months ago
README.md	activate busco in GENcontams	4 months ago

README.md

BCCM GEN-ERA tools repository

About

No description, website, or topics provided.

Readme
2 stars
3 watching
0 forks

Releases

No releases published

Packages

No packages published

Languages

Nextflow 77.0% Python 21.1%
Shell 1.9%

Figure 4. Homepage of the BCCM GEN-ERA Github.

The BCCM GEN-ERA Github thus provides a general framework for the utilization of the genomic tools developed during the project. It also provides detailed guidelines on how to use them. Each program can indeed be executed by following the step-by-step procedure and related command lines described in the "Wiki" page (https://github.com/Lcornet/GENERA/wiki). Mandatory and optional information required for correct usage is also explained. Importantly and in compliance with the FAIR approach of the project, the tools can be reproduced and redeployed by external users on their own infrastructures since all codes are shared (open source) on the "Code" window of the Github homepage.

The BCCM GEN-ERA project aimed at a long-term implementation of genomics within the collections and the acquisition of genomics skills by its members. Therefore, BCCM scientists received, in addition to the written tutorials, a dedicated training on the various software. The training session was recorded allowing future BCCM members to follow the training. The recording was also placed on the "Wiki" page of the Github as well as on Youtube (https://www.youtube.com/watch?v=bfqSl9Mi8eA) for potential external users. Moreover, access to the HPC clusters, co-financed by the BCCM GEN-ERA project and managed by the University of Liège, will be maintained after the end of the project, ensuring future utilization of the bioinformatics infrastructure.

Finally, a publication on the Github was prepared to communicate on its existence towards the scientific community. It contains additional analyses that provide a proof of concept on its usefulness and efficiency in resolving scientific questions by means of genomic analyses.

5.2 Public microbial strains and associated genome sequences

Placing microbial strains and genome sequences in the public domain is a fundamental cornerstone in science. It allows follow-on researches and the reproducibility of the analyses while avoiding duplication of efforts. Microbial strains that were investigated in the frame of the BCCM GEN-ERA project are part of the BCCM public collections. They can thus be ordered through the online catalogue and distributed to other laboratories for future studies. Similarly, whole genome sequences produced during the project were made (or will be made) publicly available on the European Nucleotide Archive (ENA). ENA is part of the International Nucleotide Sequence Database Collaboration which comprises also GenBank (United States) and the DNA DataBank of Japan. Genome sequences deposited on ENA are thus retrievable in these other repositories as well, ensuring a worldwide accessibility.

BCCM strains investigated during the project and the accession number of their genome sequence are listed in TABLE V. Access to microbial strains with whole genome sequences are increasingly requested and required. The BCCM GEN-ERA project allowed the collections to broaden their catalogue of strains with WGS data, which will thus greatly improve their visibility. More importantly, it will possible to pursue this effort in the future thanks to acquired expertise.

TABLE V. Overview of the BCCM strains with whole genome sequence acquired during the BCCM GEN-ERA project.

BCCM accession number	Species	Type of organism	ENA genome accession number
IHEM 13979	*Trichophyton kuryangei*	Fungi (mould)	GCA_910591655.1
IHEM 4712	*Trichophyton kuryangei*	Fungi (mould)	GCA_910591595.1
IHEM 13968	*Trichophyton megninii*	Fungi (mould)	GCA_910591615.1
IHEM 13976	*Trichophyton megninii*	Fungi (mould)	GCA_910591905.1
IHEM 25556	*Trichophyton rubrum*	Fungi (mould)	GCA_910591955.1
IHEM 26523	*Trichophyton rubrum*	Fungi (mould)	GCA_910592265.1
IHEM 26721	*Trichophyton rubrum*	Fungi (mould)	GCA_910592115.1
IHEM 4915	*Trichophyton rubrum*	Fungi (mould)	GCA_910591845.1
IHEM 26520	*Trichophyton rubrum*	Fungi (mould)	GCA_910592315.1
IHEM 13459	*Trichophyton soudanense*	Fungi (mould)	GCA_910591815.1
IHEM 19743	*Trichophyton soudanense*	Fungi (mould)	GCA_910592065.1
IHEM 19744	*Trichophyton soudanense*	Fungi (mould)	GCA_910592025.1
IHEM 19751	*Trichophyton soudanense*	Fungi (mould)	GCA_910592235.1
IHEM 13775	*Trichophyton violaceum*	Fungi (mould)	GCA_910591785.1
IHEM 25578	*Trichophyton violaceum*	Fungi (mould)	GCA_910592165.1
IHEM 26519	*Trichophyton violaceum*	Fungi (mould)	GCA_910592145.1
IHEM 13375	*Trichophyton yaoundei*	Fungi (mould)	GCA_910592095.1
IHEM 19041	*Trichophyton yaoundei*	Fungi (mould)	GCA_910591995.1
LMG 28623	*Apibacter mensalis*	Bacteria	To be submitted
LMG 30239	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-53654[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-53687[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-54238[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-54665[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-55228[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
R-55244[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted

R-54671[1]	*Apilactobacillus kunkeei*	Bacteria	To be submitted
LMG 30241	*Bombella apis*	Bacteria	To be submitted
R-53558[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-53711[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-53716[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-53721[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-53730[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-54679[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-54869[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-55224[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
R-55279[1]	*Bombiscardovia coagulans*	Bacteria	To be submitted
LMG 30234	*Fructobacillus fructosus*	Bacteria	To be submitted
LMG 30235	*Fructobacillus fructosus*	Bacteria	To be submitted
R-54839[1]	*Fructobacillus fructosus*	Bacteria	To be submitted
R-54866[1]	*Fructobacillus fructosus*	Bacteria	To be submitted
R-55210[1]	*Fructobacillus fructosus*	Bacteria	To be submitted
R-53534[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-53718[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-54837[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-55203[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-55214[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-55234[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
R-55250[1]	*Fructobacillus* sp. nov. 1	Bacteria	To be submitted
LMG 30237	*Fructobacillus tropaeoli*	Bacteria	To be submitted
R-53137[1]	*Fructobacillus tropaeoli*	Bacteria	To be submitted
LMG 30242	*Gilliamella bombicola*	Bacteria	To be submitted
R-53673[1]	*Hafnia alvei*	Bacteria	To be submitted
R-53703[1]	*Hafnia alvei*	Bacteria	To be submitted
R-53736[1]	*Hafnia alvei*	Bacteria	To be submitted
R-54239[1]	*Hafnia alvei*	Bacteria	To be submitted
R-54855[1]	*Hafnia alvei*	Bacteria	To be submitted
R-55238[1]	*Hafnia alvei*	Bacteria	To be submitted
R-53132[1]	*Lactobacillus bombicola*	Bacteria	To be submitted
R-53522[1]	*Lactobacillus bombicola*	Bacteria	To be submitted
R-54237[1]	*Lactobacillus bombicola*	Bacteria	To be submitted
R-55242[1]	*Lactobacillus bombicola*	Bacteria	To be submitted
R-55248[1]	*Lactobacillus bombicola*	Bacteria	To be submitted
R-53553[1]	*Lactobacillus* sp. nov. 1	Bacteria	To be submitted
R-53122[1]	*Lactococcus lactis*	Bacteria	To be submitted
R-53719[1]	*Lactococcus lactis*	Bacteria	To be submitted
R-54233[1]	*Lactococcus lactis*	Bacteria	To be submitted
R-54685[1]	*Lactococcus lactis*	Bacteria	To be submitted
R-53532[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
R-53647[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
R-53648[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
R-53650[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
R-53675[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
R-55253[1]	*Lysinibacillus fusiformis*	Bacteria	To be submitted
LMG 30244	*Secundilactobacillus kimchicus*	Bacteria	To be submitted
R-53510[1]	*Secundilactobacillus kimchicus*	Bacteria	To be submitted
R-53646[1]	*Secundilactobacillus kimchicus*	Bacteria	To be submitted
R-53584[1]	*Snodgrassella alvi*	Bacteria	To be submitted
LMG 28360	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768045
LMG 32767	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768055
LMG 32768	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768065
LMG 32769	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768085

LMG 32770	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768035
LMG 32771	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768075
LMG 32772	*Snodgrassella* sp. nov. 1: *S. communis*	Bacteria	GCA_914768015
LMG 30236	*Snodgrassella* sp. nov. 2: *S. gandavensis*	Bacteria	GCA_914768025
LMG 32773	*Snodgrassella* sp. nov. 2: *S. gandavensis*	Bacteria	GCA_914768095
MUCL 42894	*Starmerella apicola*	Fungi (yeast)	To be submitted
MUCL 53493	*Starmerella apicola*	Fungi (yeast)	To be submitted
MUCL 53486	*Starmerella apicola*	Fungi (yeast)	To be submitted
MUCL 53488	*Starmerella apicola*	Fungi (yeast)	To be submitted
MUCL 56111	*Starmerella apis*	Fungi (yeast)	To be submitted
MUCL 56112	*Starmerella apis*	Fungi (yeast)	To be submitted
MUCL 56113	*Starmerella apis*	Fungi (yeast)	To be submitted
MUCL 56105	*Starmerella bombi*	Fungi (yeast)	To be submitted
MUCL 56106	*Starmerella bombi*	Fungi (yeast)	To be submitted
MUCL 56107	*Starmerella bombi*	Fungi (yeast)	To be submitted
MUCL 42985	*Starmerella bombi*	Fungi (yeast)	To be submitted
MUCL 56084	*Starmerella bombicola*	Fungi (yeast)	To be submitted
MUCL 56086	*Starmerella bombicola*	Fungi (yeast)	To be submitted
MUCL 56100	*Starmerella magnoliae*	Fungi (yeast)	To be submitted
MUCL 56101	*Starmerella magnoliae*	Fungi (yeast)	To be submitted
MUCL 56102	*Starmerella magnoliae*	Fungi (yeast)	To be submitted
MUCL 56103	*Starmerella magnoliae*	Fungi (yeast)	To be submitted
MUCL 56083	*Starmerella magnoliae*	Fungi (yeast)	To be submitted
MUCL 45721	*Starmerella neotropicalis*	Fungi (yeast)	To be submitted
MUCL 53320	*Starmerella neotropicalis*	Fungi (yeast)	To be submitted
MUCL 52594	*Starmerella sp.*	Fungi (yeast)	To be submitted
ITM 500938	*Mycobacterium alsense*	Mycobacteria	To be submitted
ITM 500927	*Mycobacterium angelicum*	Mycobacteria	To be submitted
ITM 500012	*Mycobacterium diernhoferi*	Mycobacteria	To be submitted
ITM 500921	*Mycobacterium engbaekii*	Mycobacteria	To be submitted
ITM 500936	*Mycobacterium europaeum*	Mycobacteria	To be submitted
ITM 500939	*Mycobacterium fragae*	Mycobacteria	To be submitted
ITM 500930	*Mycobacterium heraklionense*	Mycobacteria	To be submitted
ITM 500933	*Mycobacterium iranicum*	Mycobacteria	To be submitted
ITM 500934	*Mycobacterium longobardum*	Mycobacteria	To be submitted
ITM 500034	*Mycobacterium nonchromogenicum*	Mycobacteria	To be submitted
ITM 500931	*Mycobacterium palustre*	Mycobacteria	To be submitted
ITM 500925	*Mycobacterium paraense*	Mycobacteria	To be submitted
ITM 500926	*Mycobacterium persicum*	Mycobacteria	To be submitted
ITM 500928	*Mycobacterium riyadhense*	Mycobacteria	To be submitted
ITM 500083	*Mycobacterium szulgai*	Mycobacteria	To be submitted
ITM 500028	*Mycobacterium triviale*	Mycobacteria	To be submitted
ULC 096	*Phormidium terebriform*	Cyanobacteria	To be submitted
ULC 102	*Phormidium pseudopriestleyi*	Cyanobacteria	To be submitted
ULC 722	*Oscillatoriacea* sp.	Cyanobacteria	To be submitted

[1] The R-isolates will subsequently be made publicly available in the BCCM/LMG collection.

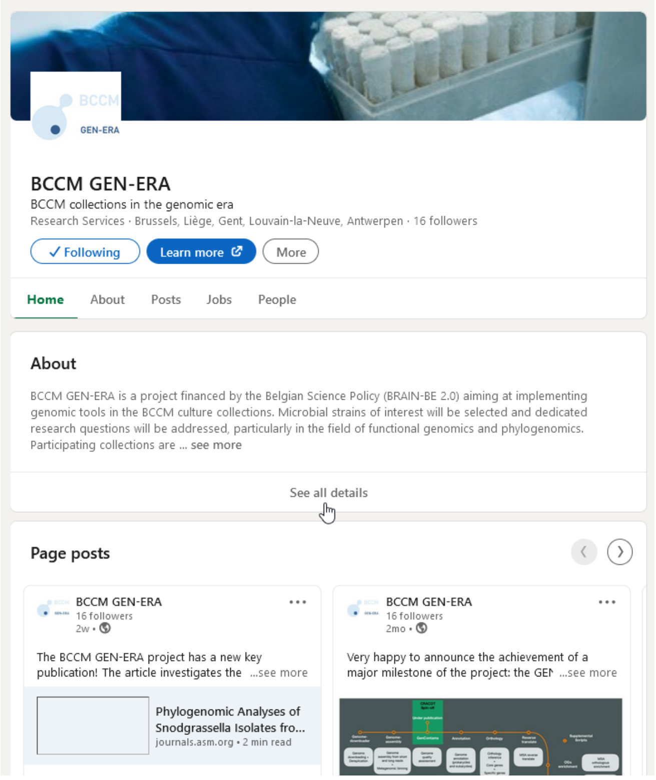
5.3 Conferences: oral and poster communications

The following list provides the oral and poster communications performed in the frame of the BCCM GEN-ERA project:

- E. D'hooge, L. Cornet. The taxonomy of the *Trichophyton rubrum* complex: a phylogenomic approach. Joint webinar of the Belgian Society for Human and Animal Mycology and the Société Française de Mycologie Médicale, 3 June 2021. Oral communication.
- P. Becker, L. Cornet, E. D'hooge, N. Magain, D. Stubbe, A. Packeu, D. Baurain. The taxonomy of the *Trichophyton rubrum* complex: a phylogenomic approach. 10th Trends in Medical Mycology, 8-11 October 2021, Aberdeen, Scotland, UK. Poster.
- L. Cornet, D. Baurain. A guided tour into genomic contamination detection. EMBL symposium "Reconstructing the human past: using ancient and modern genomics", 13-16 September 2022, Heidelberg, Germany. Poster.
- L. Cornet, D. Baurain. A guided tour into genomic contamination detection. Joint 6th Annual Meeting on Plant Ecology and Evolution & COBECORE meeting, 29-30 September 2022, Meise, Belgium. Poster.
- L. Cornet. Contamination detection in genomic data: more is not enough. Microbiome Virtual International Forum, 18 October 2022. Oral communication.

5.4 BCCM GEN-ERA Linkedin website

A LinkedIn page (https://www.linkedin.com/company/bccm-gen-era) was created for the BCCM GEN-ERA project in order to provide a general overview. Messages were regularly posted to keep the audience informed on the advancements.



BCCM GEN-ERA

BCCM collections in the genomic era

Research Services · Brussels, Liège, Gent, Louvain-la-Neuve, Antwerpen · 16 followers

About

BCCM GEN-ERA is a project financed by the Belgian Science Policy (BRAIN-BE 2.0) aiming at implementing genomic tools in the BCCM culture collections. Microbial strains of interest will be selected and dedicated research questions will be addressed, particularly in the field of functional genomics and phylogenomics. Participating collections are ... see more

Page posts

BCCM GEN-ERA
16 followers
2w •

The BCCM GEN-ERA project has a new key publication! The article investigates the ...see more

Phylogenomic Analyses of Snodgrassella Isolates fro...
journals.asm.org • 2 min read

BCCM GEN-ERA
16 followers
2mo •

Very happy to announce the achievement of a major milestone of the project: the GEN ...see more

Figure 5. Overview of the BCCM GEN-ERA LinkedIn page

5.5 New services

Following the BCCM GEN-ERA project, the participating collections were able to develop their portfolio of genomics applications. The latter include notably the provision of high-quality genomic DNA for WGS sequencing, either on BCCM strains or on isolates provided by BCCM customers. The assembly of raw reads data to deliver genome sequences as well as their annotation represent other examples of new services that BCCM collections are able to propose to its users, from both the academic and industrial sectors.

6. PUBLICATIONS

The following articles were published as open access:

- Léonard RR, Leleu M, Vlierberghe MV, Cornet L, Kerff F and Baurain D (2021). ToRQuEMaDA: tool for retrieving queried *Eubacteria*, metadata and dereplicating assemblies. PeerJ 9: e11348. https://peerj.com/articles/11348/
- Cornet L, D'hooge E, Magain N, Stubbe D, Packeu A, Baurai D and Becker P (2021). The taxonomy of the *Trichophyton rubrum* complex: a phylogenomic approach. Microbial Genomics 7: 000707. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000707
- Cornet L, Ahn A-C, Wilmotte A and Baurain D (2021). ORPER: A Workflow for Constrained SSU rRNA Phylogenies. Genes 12: 1741. https://www.mdpi.com/2073-4425/12/11/1741/html
- Cornet L, Cleenwerck I, Praet J, Leonard R, Vereecken NJ, Michez D, Smagghe G, Baurain D, Vandamme P (2022). Phylogenomic analyses of *Snodgrassella* isolates from honeybees and bumblebees reveals taxonomic and functional diversity. mSystems 7: 01500-21. https://journals.asm.org/doi/10.1128/msystems.01500-21
- Cornet L and Baurain D (2022). Contamination detection in genomic data: more is not enough. Genome Biology 23: 60. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02619-9

In addition, the following articles were submitted:

- Meunier L, Baurain D and Cornet L (2022). AMAW: automated gene annotation for non-model eukaryotic genomes. Submitted to F1000Research.
- Cornet L, Durieu B, Baert F, D'hooge E, Colignon D, Meunier L, Lupo V, Cleenwerck I, Daniel H-M, Rigouts L, Sirjacobs D, Declerck S, Vandamme P, Wilmotte A, Baurain D, Becker P (2022). The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics. Submitted to GIGAscience.
- Cornet L, Baurain D (2022). CRitical Assessment of genomic COntamination detection at multiple Taxonomic levels (CRACOT). Submitted to Genome Biology.

7. ACKNOWLEDGEMENTS

We thank the Belgian Science Policy Office for financing the BCCM GEN-ERA project.

We thank the CECI and especially David Colignon for his help in the installation of the Galaxy instance, the support for the training, and the reviewing of the GEN-ERA documentation. We also thank Olivier Mattelaer for his help with Singularity learning. For the user's connection to the ULiege VPN and for virtual machine availability, we thank the UNIPC and especially Michelle Dupont and Eric Palla. Thanks also to Damien Sirjacobs for the user management. We warmly thank Valerian Lupo for the Physeter support and Loic Meunier for the AMAW support.

Finally, we thank the members of the follow-up committee of the BCCM GEN-ERA project for their participation, opinion, help, recommendations, guidance and competent advices:

- Ir. Maria-Helena Bosschaerts (BELSPO, Belgium)
- Dr. Muriel Gugger (Pasteur Institute, Paris, France)
- Prof. Marie-Pierre Hayette (CHU de Liège, Belgium)
- Dr. Ferry Hagen (Westerdijk Fungal Biodiversity Institute, The Netherlands)
- Prof. Dick van Soolingen (RIVM, The Netherlands)
- Dr. Richard Anthony (RIVM, The Netherlands)
- Dr. Sébastien Rigali (ULiège, Belgium)
- Dr. Peter Chaerle (UGent, Belgium)

8. REFERENCES

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46: W537–544.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Loman NJ, Andersson AF, Quince C (2013). CONCOCT: Clustering contigs on coverage and composition. Available from: http://arxiv.org/abs/1312.4038.

Archibald JM (2009). The puzzle of plastid evolution. Current Biology 19: R81–R88.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455–477.

Becker P, Bosschaerts M, Chaerle P, Daniel H-M, Hellemans A, Olbrechts A, Rigouts L, Wilmotte A, Hendrickx M (2019). Public microbial resources centres: key hubs for FAIR microorganisms and genetic materials. Appl Environ Microbiol 85: e01444-19.

Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics 3: lqaa108.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36: 1925–1927.

Claus S, Van Bogaert INA (2017) Sophorolipid production by yeasts: a critical review of the literature and suggestions for future research. Appl Microbiol Biotechnol 101: 7811–7821.

Cornet L, Ahn A-C, Wilmotte A, Baurain D (2021). ORPER: A workflow for constrained SSU rRNA phylogenies. Genes 12: 1741.

Cornet L, Baurain D (2022). Contamination detection in genomic data: more is not enough. Genome Biology 23: 60.

Criscuolo A, Gribaldo S (2010). BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210.

de Hoog GS, Dukik K, Monod M, Packeu A, Stubbe D, Hendrickx M ,Kupsch C, Stielow JB, Freeke J, Göker M, Rezaei-Matehkolaei A, Mirhendi H, Gräser Y (2017). Toward a novel multilocus phylogenetic taxonomy for the dermatophytes. Mycopathol 182: 5–31.

Dessimoz C, Gil M (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biology 11: R37.

Detry R, Simon-Delso N, Bruneau E, Daniel H-M (2020). Specialisation of yeast genera in different phases of bee bread maturation. *Microorganisms 8*: 1789.

Emms DM, Kelly S (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20: 238.

Engel P, Stepanauskas R, Moran NA (2014). Hidden diversity in honeybee gut symbionts detected by single-cell genomics. PLoS Genet 10: e1004596.

Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3: e1319.

Federhen S (2012). The NCBI Taxonomy database. Nucleic Acids Research 40: D136–143.

Gonçalves C, Wisecaver JH, Kominek J, Oom MS, Leandro MJ, Shen X-X, Opulente DA, Zhou X, Peris D, Kurtzman CP, Hittinger CT, Rokas A, Gonçalves P (2018). Evidence for loss and reacquisition of alcoholic fermentation in a fructophilic yeast lineage. eLife 7: e33034.

Gonçalves P, Gonçalves C, Brito PH, Sampaio JP (2020). The *Wickerhamiella/Starmerella* clade: a treasure trove for the study of the evolution of yeast metabolism. Yeast 37: 313–320.

Gonçalves C, Marques M, Gonçalves P (2022). Contrasting strategies for sucrose utilization in a floral yeast blade. mSphere 7: e0003522.

Gontcharov AA, Marin B, Melkonian M (2004). Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). Molecular Biology and Evolution 21: 612–24.

Gräser Y, Kuijpers AF, Presber W, de Hoog GS (2000). Molecular taxonomy of the *Trichophyton rubrum* complex. J Clin Microbiol 38: 3329–3336.

Hunt M, De Silva N, Otto TD, Parkhill J, Keane JA, Harris SR (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biology 16: 294.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, Philippe H (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nature Ecology and Evolution 1: 1370–1378.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9: 5114.

Jauffrit F, Penel S, Delmotte S, Rey C, de Vienne DM, Gouy M, Charrier J-P, Flandrois J-P, Brochier-Armanet C (2016). RiboDB database: a comprehensive resource for prokaryotic systematics. Molecular Biology and Evolution 33: 2170– 2172.

Kanehisa M, Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7: e7359.

Knoll AH (2003). The geological consequences of evolution. Geobiology 1: 3–14.

Koch H, Abrol DP, Li J, Schmid-Hempel P (2013). Diversity and evolutionary patterns of bacterial gut associates of corbiculate bees. Mol Ecol 22: 2028–2044.

Kolmogorov M, Yuan J, Lin Y, Pevzner P (2019). Assembly of long error-prone reads using repeat graphs. Nat Biotechnol 37: 540-546.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods 17: 1103–1110.

Kominek J, Doering DT, Opulente DA, Shen X-X, Zhou X, DeVirgilio J, Hulfachor AB, Groenewald M, Mcgee MA, Karlen SD, Kurtzman CP, Rokas A, Hittinger CT (2019). Eukaryotic acquisition of a bacterial operon. Cell 176: 1356-1366.e10.

Kopp RE, Kirschvink JL, Hilburn IA, Nash CZ (2005). The paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. PNAS 102: 11131–11136.

Kwong WK, Medina LA, Koch H, Sing K-W, Soh EJY, Ascher JS, Jaffé R, Moran NA (2017). Dynamic microbiome evolution in social bees. Sci Adv 3: e1600513.

Kwong WK, Moran NA (2013). Cultivation and characterization of the gut symbionts of honeybees and bumble bees: description of *Snodgrassella alvi* gen. nov., sp. nov., a member of the family *Neisseriaceae* of the Beta- proteobacteria, and *Gilliamella apicola* gen. nov., sp. nov., a member of *Orbaceae* fam. nov., *Orbales* ord. nov., a sister taxon to the order "*Enterobacteriales*" of the Gammaproteobacteria. Int J Syst Evol Microbiol 63: 2008–2018.

Li L, Stoeckert CJ, Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178–2189.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Research 18: 298–309.

Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L (2021). Contamination in reference sequence databases: time for divide-and-rule tactics. Frontiers in Microbiology 12: 3233.

Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution 38: 4647–4654.

Martinson VG, Danforth BN, Minckley RL, Rueppell O, Tingek S, Moran NA (2011). A simple and distinctive microbiota associated with honeybees and bumble bees. Mol Ecol 20: 619–628.

Meunier L, Baurain D, Cornet L (2022). AMAW: automated gene annotation for non-model eukaryotic genomes. Available from: https://www.biorxiv.org/content/10.1101/2021.12.07.471566v1.

Moore KR, Magnabosco C, Momper L, Gold DA, Bosak T, Fournier GP (2019). An expanded ribosomal phylogeny of cyanobacteria supports a deep placement of plastids. Frontiers in Microbiology 10: 1612.

Nasko DJ, Koren S, Phillippy AM, Treangen TJ (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biology 19: 165.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Research 27: 824–834.

Ochoa de Alda JAG, Esteban R, Diago ML, Houmard J (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun 5: 4937.

Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology 22: 178.

Packeu A, Stubbe D, Roesems S, Goens K, Van Rooij P, de Hoog GS, Hendrickx M (2020). Lineages within the *Trichophyton rubrum* complex. Mycopathol 185: 123–136.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25: 1043–1055.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P (2018). A standardized bacterial taxonomy bases on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36: 996-1004.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P (2019). Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. Available from: https://www.biorxiv.org/content/10.1101/771964v1.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 38: 1079–86.

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research 50: D785–D794.

Persinoti GF, Martinez DA, Li W, Döğen A, Billmyre RB, Averette A, Goldberg JM, Shea T, Young S, Zeng Q, Oliver BG, Barton R, Metin B, Hilmioğlu-Polat S, Ilkit M, Gräser Y, Martinez-Rossi NM, White TC, Heitman J, Cuomo CA (2018). Whole-genome analysis illustrates global clonal population structure of the ubiquitous dermatophyte pathogen *Trichophyton rubrum*. Genetics 208: 1657-1669.

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, Cruaud C, Labadie K, Aury J-M, Istace B, Lebrigand K, Barbry P, Engelen S, Lemainque A, Wincker P, Liti G, Schacherer J (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 556: 339–344 (2018).

Praet J, Parmentier A, Schmid-Hempel R, Meeus I, Smagghe G, Vandamme P (2018). Large-scale cultivation of the bumblebee gut microbiota reveals an underestimated bacterial species diversity capable of pathogen inhibition. Environ Microbiol 20: 214–227.

Qazi MA, Wang Q, Dai Z (2022). Sophorolipids bioproduction in the yeast *Starmerella bombicola*: current trends and perspectives. Bioresource Technology 346: 126593.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596.

Queirós P, Delogu F, Hickl O, May P, Wilmes P (2021). Mantis: flexible and consensus-driven genome annotation. GigaScience 10: giab042.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RYY (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Microbiology 111: 1–61.

Rodríguez A, Burgon JD, Lyra M, Irisarri I, Baurain D, Blaustein L, Göçmen B, Künzel S, Mable BK, Nolte AW, Veith M, Steinfartz S, Elmer KR, Philippe H, Vences M (2017). Inferring the shallow phylogeny of true salamanders (Salamandra) by multiple phylogenomic approaches. Molecular Phylogenetics and Evolution 115: 16–26.

Roure B, Rodriguez-Ezpeleta N, Philippe H (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evolutionary Biology 7: S2.

Saary P, Mitchell AL, Finn RD (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. Genome Biology 21: 244.

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2022). GenBank. Nucleic Acids Research 50: D161–164.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020: baaa062.

Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068 2069.

Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe R, Cadez N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A (2018). Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175: 1533-1545.e20.

Stamatakis A, Hoover P, Rougemont J (2008). A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57: 758–771.

Steele MI, Moran NA (2021). Evolution of interbacterial antagonism in bee gut microbiota reflects host and symbiont diversification. mSystems 6: e00063-21.

Steenwyk JL, Opulente DA, Kominek J, Shen X-X, Zhou X, Labella AL, Bradley NP, Eichman BF, Cadez N, Libkind D, DeVirgilio J, Hulfachor AB, Kurtzman CP, Hittinger CT, Rokas A (2019). Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol 17: e3000255.

Su H, Packeu A, Ahmed SA, Al-Hatmi AMS, Blechert O, Ilkit M, Hagen F, Gräser Y, Liu W, Deng S, Hendrickx M, Xu J, Zhu M, de Hoog S (2019). Species distinction in the *Trichophyton rubrum* complex. J Clin Microbiol 57: e00352-19.

Tortoli E, Fedrizzi T, Meehan CJ, Trovato A, Grottola A, Giacobazzi E, Serpini GF, Tagliazucchi S, Fabio A, Bettua C, Bertorelli R, Frascaro F, De Sanctis V, Pecorari M, Jousson O, Segata N, Cirillo DM (2017). The new phylogeny of the genus *Mycobacterium*: the old and the news. Infect Genet Evol 56: 19–25.

Van Bogaert INA, Holvoet K, Roelants SLKW, Li B, Lin YC, Van de Peer Y, Soetaert W (2013). The biosynthetic gene cluster for sophorolipids: a biotechnological interesting biosurfactant produced by *Starmerella bombicola*. Mol Microbiol 88: 501–509.

Whitton BA, Potts M (2012). Introduction to the cyanobacteria. In: Whitton BA, editor. Ecology of cyanobacteria II: their diversity in space and time. Springer, Dordrecht, The Netherlands. pp. 1–13.

Wood DE, Lu J, Langmead B (2019). Improved metagenomic analysis with Kraken 2. Genome Biology 20: 257.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rossello-Mora R (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12: 635–645.

Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV (2021). OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Research 49: D389–393.

Zhan P, Dukik K, Li D, Sun J, Stielow JB, Gerrits van den Ende B, Brankovics B, Menken SBJ, Mei H, Bao W, Lv G, Liu W, de Hoog GS (2018). Phylogeny of dermatophytes with genomic character evaluation of clinically distinct *Trichophyton rubrum* and *T. violaceum*. Stud Mycol 89: 153-175.

Zhang Z-J, Huang M-F, Qiu L-F, Song R-H, Zhang Z-X, Ding Y-W, Zhou X, Zhang X, Zheng H (2021). Diversity and functional analysis of Chinese bumblebee gut microbiota reveal the metabolic niche and antibiotic resistance variation of *Gilliamella*. Insect Sci 28:302–314.