GigaScience, 2023, **12**, 1–10
DOI: 10.1093/gigascience/giad022
TECH NOTE

The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics

Luc Cornet [1,*], Benoit Durieu [2], Frederik Baert [1], Elizabet D'hooge [1], David Colignon [3], Loic Meunier [4], Valérian Lupo [4], Ilse Cleenwerck [5], Heide-Marie Daniel [6], Leen Rigouts [7], Damien Sirjacobs [4], Stéphane Declerck [6], Peter Vandamme [5], Annick Wilmotte [2], Denis Baurain [4] and Pierre Becker [1]

[1]BCCM/IHEM, Mycology and Aerobiology, Sciensano, 1050, Brussels, Belgium
[2]InBioS, Physiology and Bacterial Genetics, University of Liège, 4000, Liège, Belgium
[3]Applied and Computational Electromagnetics (ACE), University of Liège, 4000, Liège, Belgium
[4]InBioS–PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, 4000, Liège, Belgium
[5]BCCM/LMG and Laboratory of Microbiology, Faculty of Sciences, Ghent University, 9000, Ghent, Belgium
[6]BCCM/MUCL and Laboratory of Mycology, Earth and Life Institute, Université catholique de Louvain, ELIM 1348, Louvain-la-Neuve, Belgium
[7]BCCM/ITM, Mycobacteriology Unit, Institute of Tropical Medicine, 2000, Antwerp, Belgium
*Correspondence address. Luc Cornet. Luc Cornet University of Liege InBios CIP Allée du 6 Août, 114000; E-mail: luc.cornet@uliege.be

Abstract

Background: Microbial culture collections play a key role in taxonomy by studying the diversity of their strains and providing well-characterized biological material to the scientific community for fundamental and applied research. These microbial resource centers thus need to implement new standards in species delineation, including whole-genome sequencing and phylogenomics. In this context, the genomic needs of the Belgian Coordinated Collections of Microorganisms were studied, resulting in the GEN-ERA toolbox. The latter is a unified cluster of bioinformatic workflows dedicated to both bacteria and small eukaryotes (e.g., yeasts).

Findings: This public toolbox allows researchers without a specific training in bioinformatics to perform robust phylogenomic analyses. Hence, it facilitates all steps from genome downloading and quality assessment, including genomic contamination estimation, to tree reconstruction. It also offers workflows for average nucleotide identity comparisons and metabolic modeling.

Technical details: Nextflow workflows are launched by a single command and are available on the GEN-ERA GitHub repository (https://github.com/Lcornet/GENERA). All the workflows are based on Singularity containers to increase reproducibility.

Testing: The toolbox was developed for a diversity of microorganisms, including bacteria and fungi. It was further tested on an empirical dataset of 18 (meta)genomes of early branching Cyanobacteria, providing the most up-to-date phylogenomic analysis of the *Gloeobacterales* order, the first group to diverge in the evolutionary tree of Cyanobacteria.

Conclusion: The GEN-ERA toolbox can be used to infer completely reproducible comparative genomic and metabolic analyses on prokaryotes and small eukaryotes. Although designed for routine bioinformatics of culture collections, it can also be used by all researchers interested in microbial taxonomy, as exemplified by our case study on *Gloeobacterales*.

Keywords: workflow, genomics, metagenomics, phylogeny, phylogenomics, culture collections, nextflow, Singularity containers, *Gloeobacterales*, Cyanobacteria

Background

Genomics has revolutionized a number of research fields, including microbial taxonomy. Nowadays, genomes are frequently used for species delineation; the average nucleotide identity (ANI) comparisons are becoming the new gold standard for bacterial and yeast taxonomy, replacing DNA–DNA hybridization experiments [1–4]. The Genome Taxonomy Database (GTDB) project demonstrates the usefulness of this approach by providing a prokaryotic taxonomy completely based on genome sequences [5, 6]. Complementary to ANI, phylogenomics is also increasingly used to guide the taxonomy of microorganisms, notably small eukaryotes [7–9]. Phylogenomic studies are based on the analysis of hundreds to thousands of genes at once, outperforming single-gene phylogenies in terms of resolution and accuracy [10–12].

Microbial culture collections are public biological resource centers that preserve and distribute microorganisms for many purposes, such as industrial applications, quality controls, teaching activities, or scientific research at large. They also play an important role in taxonomy, either by investigating the phylogeny of their own strains or by distributing them to taxonomists [13, 14]. To enforce a correct taxonomy for their diverse microbial materials, culture collections have to integrate modern genomic practices. This task is not trivial since genomics is a rapidly changing field, and the bioinformatic pipelines are constantly evolving. For instance, the evaluation of genomic contamination has evolved a lot during the past 3 years, with 11 new algorithms published [15]. The production of genome assemblies can also require advanced metagenomic methods, depending on the axenic level of the cultures [16, 17].

In 2016, a survey designed to evaluate the bioinformatic reproducibility in science reported that 70% of researchers failed to reproduce genomic research from other scientists and that 50% failed to reproduce their own research [18]. The main source of computational irreproducibility was due to variations be-

Received: October 24, 2022. **Revised:** January 29, 2023. **Accepted:** March 14, 2023

tween operating systems and (lack of) availability of software and databases [19]. These limitations can be overcome by the use of Singularity containers, recently renamed Apptainer from the Linux foundation, that package software in a frozen computational environment [20]. Nextflow is a Singularity-aware workflow system that is well suited to address the challenge of reproducibility [19].

The availability of reproducible genomic tools for taxonomic studies is relevant for microbial collections. In this context, the needs of 5 collections belonging to the Belgian Coordinated Collections of Microorganisms (BCCM) were addressed in the framework of the Belgian Science Policy (BELSPO) GEN-ERA project [21]. The latter aimed to establish modern genomic practices for improving the taxonomy of various types of microorganisms: molds, yeasts, cyanobacteria, mycobacteria, and endosymbiotic bacteria/fungi. We report here the implementation of 13 Nextflow workflows, supported by 14 Singularity containers, which cover the most common genomic applications related to microbial taxonomy, including metabolic modeling. To our knowledge, GEN-ERA is the first unified publicly available toolbox designed for genomic studies on bacteria and small eukaryotes. It is designed to be used by microbiologists without deep knowledge of bioinformatics. Although it was initially designed for culture collections, it has indeed a broader application and can be used by any research laboratory with interest in taxonomy and comparative genomics of microorganisms.

Findings

Here, we give only an overview of the GEN-ERA toolbox (Fig. 1), while detailed descriptions are provided in the Methods section.

GEN-ERA overview

Genome-related workflows

The first 4 workflows are related to genome acquisition and annotation. The first tool, ***Genome-downloader.nf***, automatically updates a local mirror of the NCBI Taxonomy [22, 23] at each run and then downloads the genomes according to this taxonomy. The user should specify the name of the group and the taxonomic rank (e.g., "Gloeobacterales" and "order"). The specification of the taxonomic rank makes ***Genome-downloader.nf*** resilient to changes in the NCBI Taxonomy (see, e.g., [24]) that could occur in the future.

The second tool, ***Assembly.nf***, is dedicated to genome production. This workflow can assemble genomes and metagenomes from not only Illumina short reads but also PacBio or Nanopore long-read data, thanks to the use of SPAdes [25], metaSPAdes [26], and metaFlye [27]. An option for metagenomic binning, grouping contigs into individual metagenome-assembled genomes (MAGs), with MetaBAT2 [28] and CONCOCT [29], is provided too. These 2 binning algorithms are complementary, as CONCOCT is more efficient for eukaryotic data [30] while MetaBAT2 was pretrained for prokaryotic sequences [28].

The third genome-related tool, ***GENcontams.nf***, is used for the estimation of genomic contamination, completeness, and production of genome statistics. Contamination estimation (i.e., the inclusion of foreign DNA in a genome assembly) requires the use of multiple tools to recognize contaminants more accurately [15]. Indeed, some tools are dedicated to bacterial genomes (CheckM [31], GUNC [32]), others are specific to eukaryotes (EukCC [30]), and a few can work on both domains without the ability to perform interdomain detection (BUSCO [33]). In addition, Physeter [34] and Kraken 2 [35] are 2 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***. Completeness is provided by CheckM [31] for bacteria and EukCC [30] and BUSCO [33] for eukaryotes. In addition, the genome assembly quality assessment tool QUAST [36] is provided in ***GENcontams.nf*** for classical genome statistics.

The last tools of this section are related to genome annotation. The annotation (i.e., prediction of proteins) of bacterial proteins is included in the different GEN-ERA workflows (already part of GENcontams.nf, Orthology.nf, and Metabolic.nf), but we nevertheless provide a Singularity container for bacterial protein prediction with Prodigal [37]. In opposition to bacteria, eukaryotic gene annotation is not automatic in the GEN-ERA suite, but 2 tools, ***AMAW*** [38] and ***BRAKER.nf***, are included for this usage. The workflow ***BRAKER.nf*** is able to download RNA sequencing (RNA-seq) evidence, based on a user-provided list, and to use proteins from OrthoDB [39] to annotate genomes with BRAKER2 [40]. In contrast, ***AMAW*** automatizes evidence collection based on the species name [38] and is dedicated to annotation of nonmodel organisms.

Phylogeny-related workflows

This section covers phylogenomic analysis from orthology inference to production of phylogenomic trees. The first workflow, ***Orthology.nf***, implements orthology inference. Bacterial genomes (or proteomes) and eukaryotic proteomes are the basis of ***Orthology.nf.*** Two software tools can be used to compute orthologous groups (OGs) of proteins: OrthoMCL [41], available for prokaryotes only, and OrthoFinder [42], available for both domains. ***Orthology.nf*** automatically provides the core genes, shared by all genomes provided by the user in unicopy, and the specific genes, found only in a user-provided list of organisms. The OGs of proteins can be further enriched with orthologous sequences from new organisms, without running a new orthologous inference, by ***OGsEnrichment.nf***, using Forty-Two [43, 44]. OGs can also be reverse translated by ***OGsRtranslate.nf***, using Leel ([45]; available at https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo). Both protein and nucleotidic OGs can then be used for phylogenomic analysis with ***Phylogeny.nf***. This workflow implements phylogenomic inference using BMGE [46] for selection of unambiguously aligned sites, SCaFoS [47] for sequence concatenation, and RAxML [48] for tree reconstruction. With a user interface very similar to ***Phylogeny.nf***, both types of OGs can also be provided to ***PhylogenySingle.nf*** in order to compute single-gene trees with RAxML [48].

The last tool of this section is ***ORPER.nf***, which was published independently [49] and is designed to constrain a small-subunit ribosomal RNA (SSU rRNA) phylogeny with a phylogenomic backbone [49]. This tool first produces a phylogenomic tree based on concatenated ribosomal proteins, extracted from public genomes, and then constrains the larger SSU rRNA phylogeny using this reference phylogenomic tree. This multilocus constraint is used to reduce the inaccuracy of single-gene analyses [49]. ORPER permits to localize new lineages, based on SSU rRNA diversity, without a sequenced genome or to identify genomes close to strains for which only SSU rRNA sequences are available.

Other workflows

Three additional workflows are provided in the GEN-ERA toolbox. The first one, ***ANI.nf***, computes average nucleotide distances between genomes using fastANI [50]. The second one, ***GTDB.nf***, uses GTDBTk [51] for taxonomic classification of prokaryotic genomes according to the GTDB [5, 6]. The last workflow, ***Metabolic.nf***, is

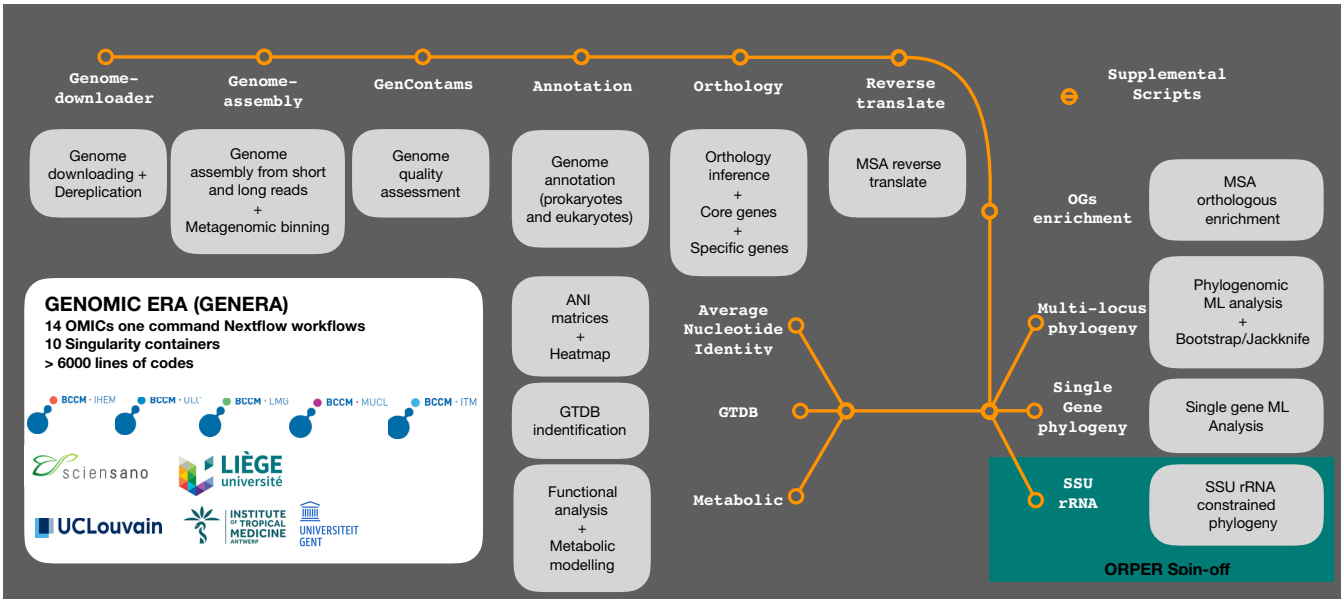

Figure 1: Overview of the GEN-ERA toolbox.

dedicated to protein function annotation using Mantis [52] and metabolic modeling of prokaryotes using Anvi'o [53] with the KEGG database as a reference [54].

Implementation

The workflows are developed with the Nextflow workflow system [19] and are all supported by Singularity containers [20]. Each workflow is accompanied by a Python script for parsing and formatting results, included in the containers. The workflows are provided to the users as programs, and each includes a help section. They can be run with a single command, increasing the reproducibility of the analyses. The databases used by the different workflows (Table 1) are automatically downloaded at the first run of the workflow if not preinstalled by the user. The GEN-ERA toolbox (workflows, Singularity definition files, companion scripts) is freely available from the GitHub repository [55]. This repository includes a detailed user guide for each tool, focusing notably on High Performance Computing (HPC) cluster usage.

Testing

The GEN-ERA toolbox was initially tested by the users from the BCCM involved in the GEN-ERA project, who were thus considered beta testers, on a SLURM-operated HPC system (durandal2/nic5, CÉCI-ULiège). These users were not advanced bioinformatics researchers, and the user guide was developed based on their needs to ensure an easy-to-use toolbox. This toolbox was further tested on the *Gloeobacterales* order (Cyanobacteria) as a case study. All command lines used for this test case are provided in Supplemental Note 1.

Gloeobacterales as a case study

Composed of thylakoid-less bacteria [56, 57], *Gloeobacterales* are the most basal order of the photosynthetic Cyanobacteria phylum. Being the first group to have diverged, it is of particular interest for the study of cyanobacterial evolution. This order has long been represented by only 2 genomes (see, e.g., Cornet et al. [58] and Moore et al. [59] phylogenies). However, the diversity of the group was recently expanded with new genomes obtained from cultivated strains [60, 61] and from metagenomes [56, 62, 63]. *Gloeobacter* spp. strains were isolated from rock biofilms, but the SSU sequences and metagenomes data show that they are widely distributed [56, 64]. For instance, the metagenomes of *Aurora vendensis* were isolated from the benthic microbial mats in an Antarctic lake [62] and the strain *Anthocerotibacter panamensis* from the surface-sterilized thallus of the hornwort *Leiosporoceros dussii* from Panama [61]. Here, we used the GEN-ERA toolbox to produce, in a completely reproducible manner, the most up-to-date phylogeny of the *Gloeobacterales* order, composed of 8 (meta)genomes (Fig. 2A, Supplemental File 1). In brief, we downloaded the genomes, estimated their contamination level, reassembled a genome deleted from the NCBI repository, and then computed large amino acid and nucleotide phylogenomic analyses, both supported by bootstrap and jackknife resampling (Fig. 2A, Supplemental File 1). Seven *Gloeobacterales* genomes were available on NCBI servers and were automatically downloaded by our tools (see Supplemental Note 1). One additional genome of *Gloeobacterales*, *Gloeobacteraceae* cyanobacterium ES-bin-313 from an Arctic glacier [63], had been deleted from NCBI servers due to a low completeness. We reassembled this genome from the raw reads and used the assembly in a phylogenomic analysis of the group for the first time. The automatization of the GEN-ERA workflows allowed us to automatically include all available strains in our phylogenies. The Supplementary Figs. S1 to S4 show 2 clusters, one with the (meta)genomes of *Gloeobacter* spp. and the other with the (meta)genomes of candidatus *A. vandensis* and *A. panamensis*, as expected [61]. We also used 566 SSU rRNA sequences from the SILVA repository [65] to estimate the sequencing level of the order (i.e., the presence and localization of the genomes among the SSU rRNA diversity) by computing an SSU rRNA phylogeny constrained by the 8 public genomes thanks to ORPER [49] (Fig. 2B). The constrained SSU rRNA phylogeny revealed 11 sequences branching at a very basal position in the cyanobacterial tree, before any known *Gloeobacterales* genomes, an observation never made before, as far as we know. These sequences likely represent interesting targets for future whole-genome sequencing projects. We also applied ANI comparisons to the 8 publicly available genomes and investigated the presence of biosynthesis KEGG pathways in *Gloeobacterales* and closely associated strains. Our results demon-

Table 1: Purpose of the GEN-ERA tools along with their databases and availability of Singularity containers

Tool	Purpose	Databases used	Availability of containers
Genome-downloader.nf	Download of NCBI genomes and proteomes	NCBI Taxonomy V: automatic setup	Yes
Assembly.nf	Assembly of (meta)genomes from short and long reads, binning of metagenomes	None	Yes
GENcontams.nf	Estimation of genome quality	NCBI Taxonomy VJune 13th 2021 GUNC: progenomes2.1 Physeter: Cornet et al., 2021 BUSCO db Vodb.10 Kraken db STD+ eukcc2_db_ver_1.1	Yes
AMAW	Eukaryotic genome annotation	prot_dbEnsembl Protists, Fungi, and Plants release 35.0 in combination with protist genomes available on the NCBI in March 2017 augustus_db VJune 28th 2021	No
Braker.nf	Eukaryotic genome annotation	OrthoDB Vodb10 Augustusdb VJune 28th 2021	No
Orthology.nf	Orthologous inference, delineation of core and specific genes	NCBI Taxonomy VJune 13th 2021	Yes
OGsEnrichment.nf	Orthologous enrichment of amino acid OGs with sequences from genomes and proteomes	NCBI Taxonomy June 13th 2021	Yes
OGsRtranslate.nf	Reverse translation of amino acid OGs	None	Yes
Phylogeny.nf	Maximum likelihood (ML) phylogenomic analysis, with bootstrap and jackknife replicates, of amino acid and nucleotide sequences	None	Yes
PhylogenySingle.nf	Single-gene ML phylogeny of amino acid and nucleotide sequences	None	Yes
ORPER.nf	SSU rRNA constrained ML phylogeny	RiboDB	Yes
ANI.nf	Average nucleotide identity comparison	None	Yes
GTDB.nf	Genome identification according to GTDB	GTDB version Vr207	Yes
Metabolic.nf	Functional and metabolic analyses	MantisDB V1.5.4 KEGG version V202	No

strate the absence of a metabolic pathway involved in the citrate cycle in the *Gloeobacterales* order (Supplemental Note 1). Two other pathways involved in carotene and isoprenoid biosynthesis are absent from the *Gloeobacter* group but present in all other sampled Cyanobacteria, with the exception of the marine *Synechococcus* sp. PCC7336 (Fig. 2C). *Anthocerotibacter panamensis* C109 is the only sampled cyanobacterium to present the archaeal (M00365) isoprenoid biosynthesis pathway (Fig. 2C). This might result from a genuine lateral gene transfer, because the contamination level of this genome is very low (0.85%).

Utilization of the GEN-ERA toolbox for the *Gloeobacterales* case study

The GEN-ERA toolbox allowed a full genomic analysis of the *Gloeobacterales* order. Although it was developed to respond to the genomics need of culture collections, this case study showed that the toolbox can be used for any comparative genomics of microorganisms, using genomic or metagenomic (public) sequencing data. Indeed, it allowed us to reassemble metagenomes and to make the binning (the latter was deleted from NCBI servers). Using the toolbox, public genomes were also downloaded and their quality estimated, notably the genomic contamination. The inference of core genes from these genomes was performed thanks to the orthologous inference and the maximum likelihood phylogenomic analyses, with bootstrap support and jackknife resampling. Constrained SSU rRNA phylogeny of the order was also inferred to provide a phylogenetic position of the sequenced organisms within the diversity represented by SSU rRNA from *Gloeobacterales*. Finally, a metabolic modeling and average nucleotide identity analyses were determined. This deep analysis of the order was performed with 10 single-command workflows, ensuring a completely reproducible study, without the need of program installation. Compared to other toolboxes, such as Atlas [66] or BACTOPIA [67], which are mainly designed for sequence analyses of bacteria, the GEN-ERA toolbox is designed for comparative genomics of both bacteria and small eukaryotes. Detailed results and examples of the practical usage of the GEN-ERA toolbox are available in Supplemental File 1.

Methods

The versions of the programs used in the case study are provided below and correspond to the first public release of the GEN-ERA (RRID:SCR_023113) toolbox (Table 1).

Genome-downloader.nf

A list of GCF accessions, from RefSeq [68, 69], and GCA accessions, from GenBank [70, 71], is created based on the assembly summary lists available on the NCBI FTP repository [23]. A local mirror of the NCBI Taxonomy is loaded with the script *setup-taxdir.pl* V0.212670 from the Bio-MUST-Core suite (available at [72]). The taxonomic lineage, from phylum to species, of each genome is obtained based on the GCF/GCA number with the companion script *fetch-tax.pl* V0.212670 (also available at [72]). Genomes are then downloaded according to the taxon name and taxonomic rank specified by the user. Priority is given to GCF over GCA assemblies for download.

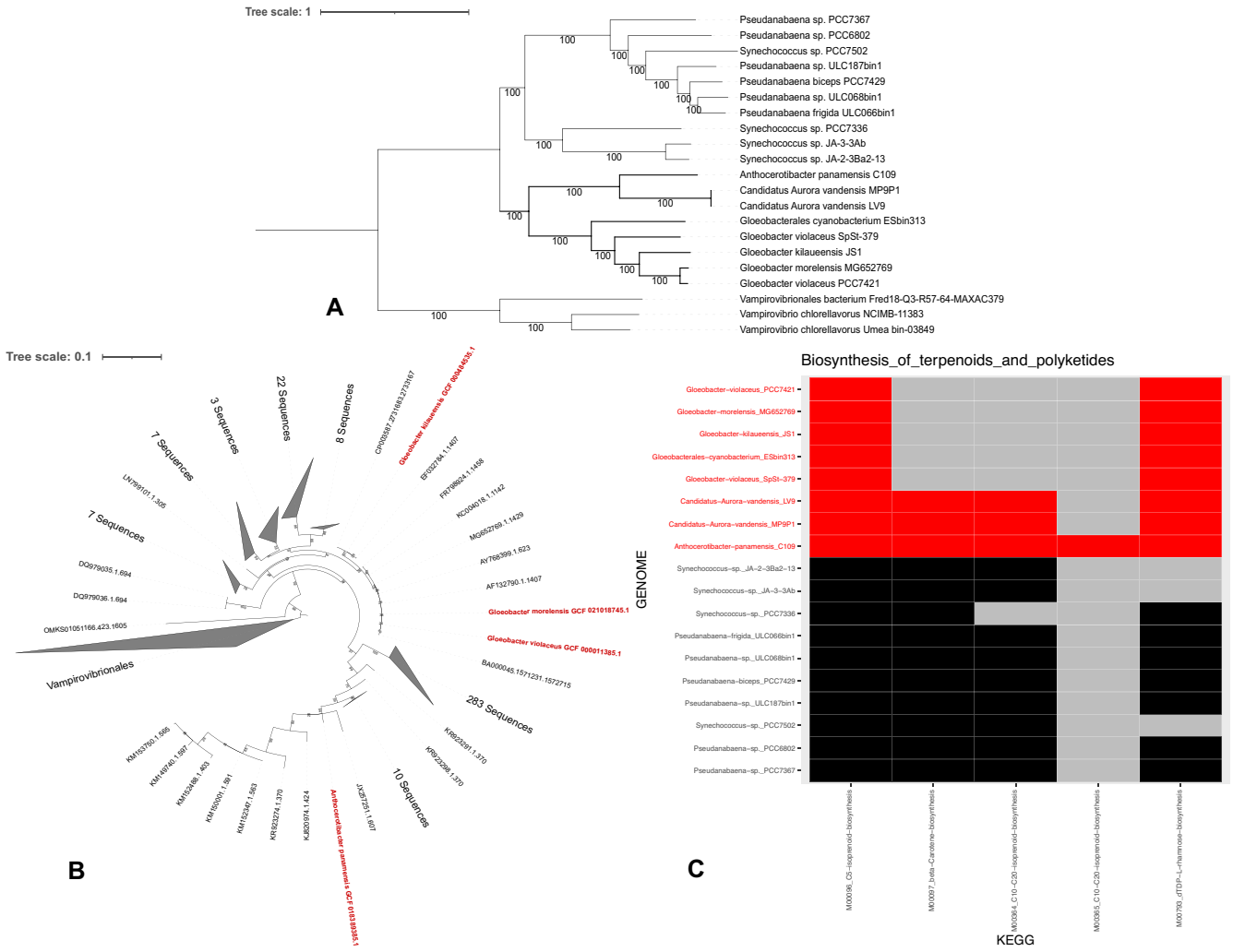

Figure 2: Results of the *Gloeobacterales* analysis. (A) Phylogenomic analysis of the *Gloeobacterales* order, conducted on 198 core genes using DNA sequences. The tree was inferred with RAxML under the GTRGAMMA model on a supermatrix of 21 × 225,524 unambiguously aligned nucleotide positions. (B) SSU rRNA phylogeny constrained by a phylogenomic analysis of ribosomal proteins, computed with ORPER. (C) Metabolic modeling of *Gloeobacterales* and closely associated taxa. Detailed methods and results of the *Gloeobacterales* analysis are available in Supplemental File 1. *Gloeobacterales* are indicated in red.

An optional dereplication of the genomes can be performed with *dRep* V3.0.0 [73] using the dereplicate option (with or without the ignoreGenomeQuality option). Finally, the proteins of the selected genomes can be downloaded if they exist on NCBI servers. Available at [55].

Assembly.nf

This workflow can take as input both short (Illumina) and long reads (PacBio and Oxford Nanopore). Short reads are first trimmed and filtered to delete low-quality reads and adapters with *fastp* (RRID:SCR_016962) V0.23.1 [74], with default settings. If only short reads are provided, the assembly is performed with *SPAdes* (RRID:SCR_000131) V3.15.3 [25] with default settings. *metaSPAdes* V3.15.3 [26] is used if the metagenome option of the workflow is specified. If long reads are provided, the assembly can be done either with *Flye* (RRID:SCR_017016) V2.19.b1774 [27], with default settings, or *CANU* (RRID:SCR_015880) V2.3 [75], with the options stopOnLowCoverage = 5 and cnsErrorRate = 0.25. *Flye* V2.19.b1774 [27], with the meta option, is the only long-read assembler available with the metagenome option. An expected genome size should be provided by the user for all long-read assemblies. The polishing of such assemblies is carried out with *pilon* (RRID:SCR_014731) V1.24 [76], with default settings, after mapping of the short reads with *bwa mem* (RRID:SCR_010910) V0.7.17 [77] and *samtools* (RRID:SCR_002105) V1.13 [78]. The metagenomic binning to obtain individual MAGs is performed with MetaBAT2 (RRID:SCR_019134) V2.15.6 [28], with default settings, and/or *CONCOCT* V1.1 [29], with default settings too. The short-read coverage is provided as input for binning after mapping with *bwa mem* V0.7.17 [77] and *samtools* V1.13 [78]. Finally, a mapping of the contigs on a reference genome, not available for metagenomes, can be performed with RagTag V2.1.0 [79]. Available at [55].

GENcontams.nf

This workflow estimates the level of genomic contamination with 6 different algorithms. The first tool is *CheckM* (RRID:SCR_016646) V1.1.3 [31], used with the lineage_wf option and the provided database. The second algorithm is *GUNC* V1.0.5 [32], with default settings, and is used with the database Progenomes 2.1 [80]. The third tool is *BUSCO* (RRID:SCR_015008) V5.3.0 [33], used in auto-lineage mode and with the provided database. The fourth tool is *Physeter* V0.213470 [34], a parser for *DIAMOND blastx* (RRID:

SCR_016071) [81] reports. *Physeter* V0.213470 is used with the autodetect option and with the database provided in Lupo et al. [34]. The fifth algorithm is *Kraken 2* (RRID:SCR_005484) V2.1.2 [35], used with default settings. The database of *Kraken 2* corresponds to the "PlusFP" database downloaded from [82]. The sixth algorithm is EukCC [30], used with default settings and the provided database. Finally, statistics on the quality of genome assemblies are computed with *QUAST* (RRID:SCR_001228) V5.1.orc1 [36], with default settings. All the algorithms can be run independently but can also be used in one go to generate a summary table. The various databases of the different tools are automatically downloaded if not provided by the user. Available at [55].

BRAKER.nf

Eukaryotic genome annotation can be performed with *AMAW* [38], a MAKER2 (RRID:SCR_005309) [83] pipeline wrapper dedicated to nonmodel organisms and automating the orchestration of its internal annotation steps, as well as the collection of species-specific transcripts and phylogenetically related protein evidence data. *BRAKER 2* (RRID:SCR_018964) V2.1.6 [40] can also be used on eukaryotic genomes. Based on a user-provided list of RNA-seq SRA numbers, the generation of transcript hints is performed by mapping the reads using *HISAT2* (RRID:SCR_015530) V9.2.1 [84] and *samtools* V1.13 [78], with default settings. Genomes of the OrthoDB [39] repository are used as protein evidence and are available in 3 different batches: fungi, protozoa, and plants. Available at [55].

Orthology.nf

Orthology inference can be performed with OrthoFinder (RRID:SCR_017118) V2.5.4 [42], used with default settings, or with OrthoMCL (RRID:SCR_007839) [41] through the pangenomic pipeline of Anvi'o (RRID:SCR_021802) V7.1 [53]. The Anvi'o mode, available for prokaryotes only, requires the use of 9 different scripts: anvi-script-reformat-fasta (with the options simplify-names and seq-type set to NT), anvi-gen-contigs-database (with default settings), anvi-run-ncbi-cogs (with default settings), anvi-gen-genomes-storage (with default settings), anvi-pan-genome (with the options mcl-inflation set to 10 and min-occurrence set to 2), anvi-get-sequences-for-gene-clusters (with default settings), anvi-script-add-default-collection (with default settings), anvi-summarize (with default settings), and anvi-compute-gene-cluster-homogeneity (with default settings). Orthology inference usually starts from complete proteomes. Nevertheless, prokaryotic genomes can be used, as prediction for prokaryotes with prodigal (RRID:SCR_011936) [37] is included in the workflow. In contrast, eukaryotic proteins should be provided by the user to Orthology.nf. After orthology inference, Orthology.nf can compute (optional) core genes. Core genes are considered here as unicopy genes shared by all organisms (and only these organisms) of a user-specified list, without exception. Another option allows the user to determine the specific genes, considered here as genes specific to a sublist of organisms, without intruders. The main difference with core genes is that specific candidate OGs will undergo an orthologous enrichment by mining the genomes of all the organisms of the orthologous inference. This strategy is used in our analyses of the *Snodgrassella*-specific gene content [85] to prevent any orthologous delineation bias. Orthologous enrichment is performed with *Forty-Two* V0.212670 [43, 44], with the same settings as ***OGsEnrichment.nf***. Available at [55].

OGsEnrichment.nf

This workflow can take as input amino acid OGs, as produced by ***Orthology.nf***. OGs can be aligned with *MUSCLE* (RRID:SCR_011812) V3.8.31 [86], with default values. The enriching sequences can come from genomes or proteomes. In both cases, BLAST banks are built with *makeblastdb* V2.10.0 [87]. The orthologous enrichment is performed with *Forty-Two* V0.212670 [43, 44]. *Forty-Two* V0.212670 is used with a BLAST e-value of 1e-05, a max_target_seqs of 10,000, the templates_seg option set to no, the ref_org_mul set to 0.3, the ref_score_mul set to 0.99, the trim_homologues option set to on, the ali_keep_lengthened_seqs option set to keep, and the ref_brh enabled. The default aligner is *BLAST* (RRID:SCR_004870) V2.10.0, but the user can also use *exonerate* V2.2.0. Available at [55].

OGsRtranslate.nf

As for ***OGsEnrichment.nf***, OGs can be aligned with *MUSCLE* V3.8.31 [86], with default values. Protein sequence alignments are back-translated by capturing and aligning the corresponding DNA sequences with the program *Leel* V0.212670 [45] (available at [72]). Available at [55].

Multilocus phylogeny.nf

This workflow takes as input OGs produced by ***Orthology.nf***, ***OGsEnrichment.nf***, or ***OGsRtranslate.nf***. The OGs can thus contain amino acid or nucleotide sequences. As for the previous workflows, amino acid OGs can be aligned with *MUSCLE* V3.8.31 [86], with default values. Nucleotide OGs are not aligned, as they are obtained by back-translating amino acid alignments with ***OGsRtranslate.nf.*** Unambiguously aligned positions in amino acid OGs are selected with BMGE V1.12 [46], used with a "medium" mask, as specified in Bio-MUST-Core V0.212670 [72]. This selection is not performed on nucleotide OGs in order to preserve the codon phase. OGs are concatenated using *SCaFoS* V1.25 [47], with default settings. Finally, trees are inferred using *RAxML* (RRID:SCR_006086) V8.2.12 [48] with 100 bootstrap replicates under the PROTGAMMALGF model for proteins and the GTRGAMMA model for DNA sequences. DNA trees are computed either without a codon partition or with a separate partition on the third codon position or based only on the 2 first positions. Beside these large phylogenomic analyses, the workflow also computes jackknife analyses. A hundred jackknife matrices are generated with the script *jack-ali-dir.pl* V0.212670 from Bio-MUST-Core [72], using a width of 100,000 positions (modifiable by the user), and concatenated with *SCaFoS* V1.25 [47], as above. The trees are computed with *RAxML* V8.2.12 [48], as above (including codon partitions), but under the fast mode. The consensus trees, from the 100 trees obtained on the matrices, are produced with consense from the PHYLIP package V3.695 [88], used with default settings. Available at [55]. Two other workflows for phylogenetic analyses are available in the GEN-ERA toolbox: ***PhylogenySingle.nf*** and ***ORPER.nf***. ***PhylogenySingle.nf*** is a simpler version of ***Phylogeny.nf***, with the same alignment, filtering of unambiguous aligned positions, and tree reconstruction settings, but for single-gene analyses. Available at [55]. ***ORPER.nf***, designed for constrained SSU rRNA phylogenetic inference, has already been published separately [49].

ANI.nf

ANI.nf performs pairwise average nucleotide identity comparisons using *fastANI* (RRID:SCR_021091) V1.33 [50] in an all-versus-all mode, with default settings. A heatmap is then computed, according to a user-specified list of genomes, with *ggplot2* [89]. Available at [55].

GTDB.nf

This workflow allows the identification of genomes according to the GTDB taxonomy [5, 6]. ***GTDB.nf*** uses *GTDBTk* V2.2.0-r207 [51] using the classify_wf workflow, with default settings. Available at [55].

Metabolic.nf

Metabolic.nf is the last workflow of the GEN-ERA toolbox. It has 2 modes: functional or modeling. The functional mode carries out a functional characterization of protein sequences using Mantis (RRID:SCR_021001) V1.5.4 [52], with default settings, whereas the modeling mode provides modeling of KEGG pathways [54], based on the presence of at least 60% of the genes involved in a pathway, for prokaryotic genomes. This mode uses the *anvi-estimate-metabolism* of Anvi'o V7.1 [53]. Presence/absence plots of KEGG pathways are then graphically represented with *ggplot2* [89], according to a user-specified list of genomes. Available at [55].

Gloeobacterales case study

Vampirovibrionales, *Pseudanabaena*, *Synechococcus*, and *Gloeobacterales* genomes were downloaded using Genome-downloader.nf V1.0.0, with default options. The genome of *Gloeobacter violaceus* SpSt-379 has been recovered using Assembly.nf V1.0.0, on the SRA SRR7539891, with the metagenome option activated and the binner option settled to all. The cyanobacteria bins were selected using GTDB.nf V1.0.0 with default options. Genomes and bins quality were estimated using GENcontam.nf V2.0.0 with CheckM [31], GUNC [32], Kraken 2 [35], and Physeter [34] with the taxonomic level option settled to phylum. The core genes were inferred using Orthology.nf V2.0.6, on 20 public genomes, with the anvio [53] option activated. The outgroup of the analysis (*Vampirovibrionales*) was not included in the definition of the core genes (presence authorized but not mandatory). *G. violaceus* SpSt-379 (CONCOCT-bin1) was further added to the core genes using OGsEnrichment.nf V1.0.0 with blast as ftaligner option. Core genes were back translated to DNA using OGsRtranslate.nf V1.0.0 with default options. Phylogenomic analysis was performed using Phylogeny.nf V1.0.3, with the jackknife option activated, with a width option settled to 50,000 for protein and 80,000 for DNA. The constrained SSU rRNA phylogeny was inferred using ORPER.nf V1.0.0 with *Gloeobacterales* as the reference group and *Vampirovibrionales* as the outgroup. The *Gloeobacterales* SSU rRNA sequences from the SILVA [65] repository were provided. The average nucleotide identity was done using ANI.nf V1.1.0, with default options. The Metabolic modeling was inferred using Metabolic.nf V1.0.0 with default options. The command lines used for this case study are available in Supplemental File 1.

Additional Files

GENERA_Supplemental-file1.pdf

Availability of Supporting Source Code and Requirements

- Project name: GEN-ERA
- Project homepage: https://github.com/Lcornet/GENERA
- License: GNU General Public license 3 (GPL-3.0)
- RRID: SCR_023,114
- Biotools: biotools:gen-era_toolbox
- workflowhub.eu: https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.416.1
- Operating system(s): Platform independent, Singularity containers
- Programming language: Nextflow and Python
- Other requirements: None

Data availability

The data used for *Gloeobacterales* analysis were downloaded from the NCBI SRA repository (SRR7539891, SRR12931219, SRR12931218). All supporting data and materials are available in the *GigaScience* GigaDB database [55].

Abbreviations

ANI: average nucleotide identity; BCCM: Belgian Coordinated Collections of Microorganisms; BLAST: Basic Local Alignment Search Tool; BUSCO: Benchmarking Universal Single-Copy Orthologs; GTDB: Genome Taxonomy Database; KEGG: Kyoto Encyclopedia of Genes and Genomes; MAG: metagenome-assembled genome; ML: maximum likelihood; NCBI: The National Center for Biotechnology Information; OG: orthologous group; RNA-seq: RNA sequencing; SSU rRNA: small-subunit ribosomal RNA.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by a research grant (no. B2/191/P2/BCCM GEN-ERA) financed by the Belgian State—Federal Public Planning Science Policy Office (BELSPO). H.-M.D. is supported by the Belgian Science Policy Office (BELSPO) grant C5/00/BCCM. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-Le Fonds National de la Recherche Scientifique (FNRS) (2.5020.11), and through 2 research grants to D.B.: B2/191/P2/BCCM GEN-ERA (Belgian Science Policy Office—BELSPO) and CDR J.0008.20 (F.R.S.-FNRS). A.W. is senior research associate of the FRS-FNRS.

Authors' contributions

L.C., D.B., and P.B. conceived the study. L.C. developed the Nextflow workflows and Singularity containers with the help of D.C. L.M. developed AMAW. V.L. developed Physeter. L.C., B.D., F.B., and E.D. tested the workflows. L.C. ran *Gloeobacterales* analyses and drew the figures. L.C., D.B., and P.B. wrote the manuscript with the help of D.S., L.R., I.C., H.-M.D., A.W., S.D., and P.V.

Acknowledgments

We thank Olivier Mattelaer for his help with Singularity containers.

References

1. Goris, J, Konstantinidis, KT, Klappenbach, JA, *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;**57**:81–91.
2. Richter, M, Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 2009;**106**:19126–31.

3. Tindall, BJ, Rosselló-Móra, R, Busse, H-J, *et al.*. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 2010;**60**:249–66.
4. Lachance, M-A, Lee, DK, Hsiang, T. Delineating yeast species with genome average nucleotide identity: a calibration of ANI with haplontic, heterothallic metschnikowia species. *Antonie Van Leeuwenhoek* 2020;**113**:2097–106.
5. Parks, DH, Chuvochina, M, Chaumeil, P-A, *et al.* Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. *Biorxiv* 2019. bioRxiv.https://doi.org/10.1101/771964
6. Parks, DH, Chuvochina, M, Chaumeil, P-A, *et al.* A complete domain-to-species taxonomy for bacteria and archaea. *Nat Biotechnol* 2020;**38**:1079–86.
7. Cornet, L, D'hooge, E, Magain, N, *et al.* The taxonomy of the trichophyton rubrum complex: a phylogenomic approach. *Microbial Genomics* 2021;**7**. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000707
8. Galindo, LJ, López-García, P, Torruella, G, *et al.* Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. *Nat Commun* 2021;**12**:4973. doi: 10.1038/s41467-021-25308-w.
9. Keeling, PJ, Luker, MA, Palmer, JD. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol Biol Evol* 2000;**17**:23–31.
10. Dessimoz, C, Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol* 2010;**11**:R37.
11. Lunter, G, Rocco, A, Mimouni, N, *et al.* Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res* 2008;**18**:298–309.
12. Wong, KM, Suchard, MA, Huelsenbeck, JP. Alignment uncertainty and genomic analysis. *Science* 2008;**319**:473–6.
13. Smith, D. Culture collections over the world. *Int Microbiol* 2003;**6**:95–100.
14. Becker, P, Bosschaerts, M, Chaerle, P, *et al.* Public microbial resource centers: key hubs for findable, accessible, interoperable, and reusable (FAIR) microorganisms and genetic materials. *Appl Environ Microbiol American Society for Microbiology* 2019;**21**. e01444–19. doi: 10.1128/AEM.01444-19.
15. Cornet, L, Baurain, D. Contamination detection in genomic data: more is not enough. *Genome Biol* 2022;**23**:60. doi: 10.1186/s13059-022-02619-9.
16. Cornet, L, Meunier, L, Vlierberghe, MV, *et al.* Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS One* 2018;**13**:e0200323.
17. Chen, L-X, Anantharaman, K, Shaiber, A, *et al.* Accurate and complete genomes from metagenomes. *Genome Res* 2020;**30**:315–33.
18. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;**533**:452–4.
19. Di Tommaso, P, Chatzou, M, Floden, EW, *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9.
20. Kurtzer, GM, Sochat, V, Bauer, MW. Singularity: scientific containers for mobility of compute. *PLoS One* 2017;**12**: e0177459.
21. Becker, P, Cornet, L, D'hooge, E, *et al. BCCM collections in the genomic era. Final report*. 2022. 2022–40. Belgian Science Policy Office. https://www.belspo.be/belspo/brain2-be/projects/FinalReports/BCCMGENERA_FinRep.pdf
22. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* 2012;**40**:D136–43.
23. Schoch, CL, Ciufo, S, Domrachev, M, *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020;doi: 10.1093/database/baaa062
24. NCBI. NCBI Taxonomy to include phylum rank in taxonomic names. *NCBI Insights*. 2021. https://ncbiinsights.ncbi.nlm.nih.gov/2021/12/10/ncbi-taxonomy-prokaryote-phyla-added/. [Accessed 2023 Mar 8].
25. Bankevich, A, Nurk, S, Antipov, D, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.
26. Nurk, S, Meleshko, D, Korobeynikov, A, *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34.
27. Kolmogorov, M, Bickhart, DM, Behsaz, B, *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10.
28. Kang, DD, Li, F, Kirton, E, *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359.
29. Alneberg, J, Bjarnason, BS, de Bruijn, I, *et al.* Binning metagenomic contigs by coverage and composition.. Nat Methods.. **11** 1144–6 2014.doi: 10.1038/nmeth.3103.
30. Saary, P, Mitchell, AL, Finn, RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol* 2020;**21**.doi: 10.1186/s13059-020-02155-4.
31. Parks, DH, Imelfort, M, Skennerton, CT, *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55.
32. Orakov, A, Fullam, A, Coelho, LP, *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* 2021;**22**.doi: 10.1186/s13059-021-02393-0.
33. Manni, M, Berkeley, MR, Seppey, M, *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**:4647–54.
34. Lupo, V, Van Vlierberghe, M, Vanderschuren, H, *et al.* Contamination in reference sequence databases: time for divide-and-rule tactics. *Front Microbiol* 2021;**12**.doi: 10.3389/fmicb.2021.755101.
35. Wood, DE, Lu, J, Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**.https://doi.org/10.1186/s13059-019-1891-0
36. Gurevich, A, Saveliev, V, Vyahhi, N, *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5.
37. Hyatt, D, Chen, G-L, LoCascio, PF, *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 2010;**11**. https://doi.org/10.1186/1471-2105-11-119
38. Meunier, L, Baurain, D, Cornet, L. AMAW: automated gene annotation for non-model eukaryotic genomes [version 1; peer review: awaiting peer review]. F1000 Research 2023. https://doi.org/10.12688/f1000research.129161.1
39. Zdobnov, EM, Kuznetsov, D, Tegenfeldt, F, *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2021;**49**:D389–93.
40. Brůna, T, Hoff, KJ, Lomsadze, A, *et al.* BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinformatics* 2021;**3**. https://doi.org/10.1093/nargab/lqaa108
41. Li, L, Stoeckert, CJ, Roos, DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178–89.
42. Emms, DM, Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;**20**.https://doi.org/10.1186/s13059-019-1832-y

43. Irisarri, I, Baurain, D, Brinkmann, H, *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* 2017. https://doi.org/10.1038/s41559-017-0240-5
44. Simion, P, Philippe, H, Baurain, D, *et al.* A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* 2017;**27**:958–67.
45. Rodríguez, A, Burgon, JD, Lyra, M, *et al.* Inferring the shallow phylogeny of true salamanders (Salamandra) by multiple phylogenomic approaches. *Mol Phylogenet Evol* 2017;**115**:16–26.
46. Criscuolo, A, Gribaldo, S. BMGE(Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 2010;**10**:210.
47. Roure, B, Rodriguez-Ezpeleta, N, Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol* 2007;**7**. https://doi.org/10.1186/1471-2148-7-S1-S2
48. Stamatakis, A, Hoover, P, Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* 2008;**57**:758–71.
49. Cornet, L, Ahn, A-C, Wilmotte, A, *et al.* ORPER: a workflow for constrained SSU rRNA phylogenies. *Genes* 2021;**12**:1741.
50. Jain, C, Rodriguez-R, LM, Phillippy, AM, *et al.*. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**. https://doi.org/10.1038/s41467-018-07641-9
51. Chaumeil, P-A, Mussig, AJ, Hugenholtz, P, *et al.* GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database. *Bioinformatics* 2022;**38**:5315–6.
52. Queirós, P, Delogu, F, Hickl, O, *et al.* Mantis: flexible and consensus-driven genome annotation. *GigaScience* 2021;**10**.https://doi.org/10.1093/gigascience/giab042
53. Eren, AM, Esen, ÖC, Quince, C, *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015;**3**:e1319.
54. Kanehisa, M, Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.
55. Cornet, L, Durieu, B, Baert, F, *et al.*. Supporting data for "The GEN-ERA Toolbox: Unified and Reproducible Workflows for Research in Microbial Genomics." *GigaScience Database*. 2023.http://dx.doi.org/10.5524/102369
56. Grettenberger, CL. Novel Gloeobacterales spp. from diverse environments across the globe. *mSphere* 2021;**6**.doi: 10.1128/mSphere.00061-21
57. Nakamura, Y, Kaneko, T, Sato, S, *et al.* Complete genome structure of gloeobacter violaceus PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* 2003;**10**:137–45.
58. Cornet, L, Bertrand, AR, Hanikenne, M, *et al.* Metagenomic assembly of new (sub)polar cyanobacteria and their associated microbiome from non-axenic cultures. *Microbial Genomics* 2018;**4**. doi: 10.1099/mgen.0.000212.
59. Moore, KR, Magnabosco, C, Momper, L, *et al.* An expanded ribosomal phylogeny of cyanobacteria supports a deep placement of plastids. *Front Microbiol* 2019;**10**:1612. doi: 10.3389/fmicb.2019.01612
60. Saw, JH, Cardona, T, Montejano, G. Complete genome sequencing of a novel gloeobacter species from a waterfall cave in Mexico. *Genome Biol Evol* 2021;**13**. https://doi.org/10.1093/gbe/evab264
61. Rahmatpour, N, Hauser, DA, Nelson, JM, *et al.* A novel thylakoidless isolate fills a billion-year gap in the evolution of cyanobacteria. *Curr Biol* 2021;**31**:2857–67.e4.
62. Grettenberger, CL, Sumner, DY, Wall, K, *et al.* A phylogenetically novel cyanobacterium most closely related to Gloeobacter. *ISME J* 2020;**14**:2142–52.
63. Zeng, Y, Chen, X, Madsen, AM, *et al.* Potential rhodopsin- and bacteriochlorophyll-based dual phototrophy in a high Arctic glacier. *mBio* 2020;**11**.
64. Mareš, J, Hrouzek, P, Kaňa, R, *et al.* The primitive thylakoid-less cyanobacterium gloeobacter is a common rock-dwelling organism. *PLoS One* 2013;**8**:e66323.
65. Quast, C, Pruesse, E, Yilmaz, P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.
66. Kieser, S, Brown, J, Zdobnov, EM, *et al.* ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinf* 2020;**21**. https://doi.org/10.1186/s12859-020-03585-4
67. Petit, RA, Read, TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. 2020;**5**(4), e00190–20.doi: 10.1128/mSystems.00190-20.
68. Pruitt, KD, Tatusova, T, Maglott, DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
69. O'Leary, NA, Wright, MW, Brister, JR, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45.
70. Sayers, EW, Cavanaugh, M, Clark, K, *et al.* GenBank. *Nucleic Acids Res* 2022;**50**:D161–4.
71. Clark, K, Karsch-Mizrachi, I, Lipman, DJ, *et al.*. GenBank. *Nucleic Acids Res* 2016;**44**:D67–72.
72. Denis Baurain: Bio-MUST-Core-0.212670—Core classes and utilities for Bio::MUST—metacpan.org. https://metacpan.org/dist/Bio-MUST-Core. [Last Accessed 2023 Mar 8].
73. Olm, MR, Brown, CT, Brooks, B, *et al.* dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;**11**:2864–8.
74. Chen, S, Zhou, Y, Chen, Y, *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**: i884–90.
75. Koren, S, Walenz, BP, Berlin, K, *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**: 722–36.
76. Walker, BJ, Abeel, T, Shea, T, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**: e112963.
77. Wang, MH, Cordell, HJ, Van Steen, K. Statistical methods for genome-wide association studies. *Semin Cancer Biol* 2019;**55**:53–60.
78. Li, H, Handsaker, B, Wysoker, A, *et al.* The sequence alignment/map format and samtools. *Bioinformatics* 2009;**25**: 2078–9.
79. Alonge, M, Soyk, S, Ramakrishnan, S, *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* 2019;**20**.
80. Mende, DR, Letunic, I, Maistrenko, OM, *et al.* proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res* 2020. https://doi.org/10.1093/nar/gkz1002
81. Buchfink, B, Xie, C, Huson, DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**: 59–60.
82. Ben, L. Kraken2 & Brachen databases. 2022 https://benlangmead.github.io/aws-indexes/k2

83. Holt, C, Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf* 2011;**12**.
84. Kim, D, Paggi, JM, Park, C, *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15.
85. Cornet, L, Cleenwerck, I, Praet, J, *et al*. Phylogenomic analyses of snodgrassella isolates from honeybees and bumblebees reveals taxonomic and functional diversity. *Msystems* 2022;**7**.
86. Edgar, RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf* 2004;**5**: 113.
87. Edgar, RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.
88. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. 2004. http://www.dbbm.fiocruz.br/molbiol/main.html. [Last Accessed July 2, 2022].
89. Wickham, H. *ggplot2*, *Use R!* Cham, Switzerland: Springer International Publishing; 2016.