Computational and Structural Biotechnology Journal 19 (2021) 5504–5509

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

MOSGA 2: Comparative genomics and validation tools

Roman Martin [a], Hagen Dreßler [a], Georges Hattab [a], Thomas Hackl [b], Matthias G. Fischer [b], Dominik Heider [a,*]

[a] Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Straße 6, Marburg 35043, Germany
[b] Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, Jahnstraße 29, Heidelberg 69120, Germany

ARTICLE INFO

Article history:
Received 10 August 2021
Received in revised form 23 September 2021
Accepted 24 September 2021
Available online 28 September 2021

Keywords:
Genome annotation
Comparative genomics
Phylogenetics
Quality control
Framework
Pipeline
Workflow

ABSTRACT

Due to the highly growing number of available genomic information, the need for accessible and easy-to-use analysis tools is increasing. To facilitate eukaryotic genome annotations, we created MOSGA. In this work, we show how MOSGA 2 is developed by including several advanced analyses for genomic data. Since the genomic data quality greatly impacts the annotation quality, we included multiple tools to validate and ensure high-quality user-submitted genome assemblies. Moreover, thanks to the integration of comparative genomics methods, users can benefit from a broader genomic view by analyzing multiple genomic data sets simultaneously. Further, we demonstrate the new functionalities of MOSGA 2 by different use-cases and practical examples. MOSGA 2 extends the already established application to the quality control of the genomic data and integrates and analyzes multiple genomes in a larger context, e.g., by phylogenetics.

1. Introduction

The constantly increasing number of available whole-genome sequences, mainly derived by high-throughput sequencing, provides more and more biological insights [32]. In particular, the sequencing of new bacterial, archaeal, and viral genomes is now routine practice. It has accelerated discoveries in microbial diversity and evolution, providing new insight into microbiome function, human health, and biogeochemical cycling [1,24,2,34,46]. However, the generation of high-quality eukaryotic genomes has been hampered in the past by sequencing costs and assembly complexity, limiting sequenced eukaryotes to those of medical or economical interest [7]. More recently, advances in Illumina high-throughput sequencing and emerging long-read sequencing technologies developed by Pacific Biosciences and Oxford Nanopore have also rendered the generation of genome assemblies of large eukaryotes a routine task [40,25,44]. These proceeding trends in data availability, advanced techniques, and improved affordability require more facilitated access for scientists to analyze these sequence data. To address the challenging issue of correctly annotating new genomes, we developed the Modular Open-Source Genome Annotator (MOSGA) [21]. While the number of available genomic data increases continuously, the quality of these data is as important as their availability. Low-quality genome assemblies will introduce errors [31] and, as such, affect the annotation quality. In addition, genome assemblies can be incomplete or contaminated, which does not necessarily affect the annotation process but will hinder the submission process and may lead to wrong conclusions. This study presents the new functionalities in MOSGA 2, such as new workflows for comparative genomics and the introduction of validation tools into the framework. In brief, MOSGA 2 incorporates comparative genomic workflows derived from Hackl *et al.* [9] and validation tools from Pirovano *et al.* [28].

We initially developed MOSGA to facilitate genome annotation, including an easy-to-use web interface based on a robust, scalable, modular, and reproducible platform. Besides new genome annotations, genome assemblies or multiple genomes can be placed in a bigger context like comparative genomics. Therefore, we integrated several tools into MOSGA 2 to calculate and output the phylogenetic trees, the average nucleotide identity, the completeness of the genomes, and the protein-coding gene comparison based on previously performed protein-coding gene annotations. In addition, we incorporated tools to identify the occurrence of contamination and present a new scaffold-detection method for organellar DNA sequences present in genome assemblies.

* Corresponding author.
E-mail address: dominik.heider@uni-marburg.de (D. Heider).

https://doi.org/10.1016/j.csbj.2021.09.024

2. Software changes

2.1. Phylogenetics

MOSGA 2 integrates a workflow based on nine established tools to preprocess and compute phylogenetic trees on multiple genomes. Preprocessing steps include selecting a database for genes, constructing and trimming multiple sequence alignment (MSA).

Therefore, we included BUSCO [35,43] and EukCC [30] to identify single-copy genes in genomes for the phylogenetic computation. The user has to choose from one of these tools or even combine both as potential phylogenetic markers. MOSGA 2 concatenates the chosen gene marker sequences into a new FASTA file. The user has to select a program for the MSA construction, such as MAFFT [12,13], ClustalW [42,18], or MUSCLE [8]. For an optional MSA trimming, we integrated trimAl [5] and ClipKIT [38]. Phylogenetic relationships are then reconstructed by maximum likelihood or distance algorithms using RAxML [36] or FastME [19], respectively, and the resulting tree are visualized by ggtree [45] and ape [26]. Tree rooting is optional and can be defined by marking an uploaded genome assembly as an outgroup.

2.2. Average nucleotide identity

For whole-genome similarity metrics, MOSGA 2 includes a comparative genomics workflow that calculates the Average Nucleotide Identity (ANI) across all genomes. To achieve this, we integrated FastANI [11], which compares the genomes against each other. Further, the ANI values are represented as a heatmap.

2.3. Protein-coding genes comparison

MOSGA 2 integrates a comparative genomics workflow that compares protein-coding genes from all uploaded genomes against each other. This comparison requires a previous annotation of protein-coding genes, which can be achieved via the annotation pipeline or by importing already annotated genomes as GBFF (GenBank flat format) files. This allows, for example, a comparison between different gene prediction tools or between reference and experimental annotations. Technically, MOSGA 2 extracts the protein-coding sequences and matches them against each other. Matches above a defined threshold will be binned back to a genome, and the average coding content similarities between the genomes are displayed as a heatmap. This analysis allows consistency checks for gene predictions across different genomes. This method compares the nucleotide sequence of protein-coding genes and has similarities with the concept of Average Amino Acids Identity.

2.4. Genome completeness

The identification of single-copy genes is a crucial step for phylogenetic analysis. Therefore, we integrated BUSCO and EukCC into MOSGA 2. While BUSCO's data source is OrthoDB [16], EukCC relies on PANTHER [22]. These tools can estimate the completeness of assemblies, and MOSGA 2 integrates them for validation into the annotation and the comparative genomics workflow. Genome completeness results for each genome are visualized together in the comparative genomics workflow and the annotation workflow separately.

2.5. Contamination detection

To detect potential contamination in a genome assembly, such as sequences from other organisms or residual sequencing adapters, we integrated two validation tools to identify putative contamination. In the case of a gene prediction workflow with RNA-seq based gene prediction, MOSGA 2 offers BlobTools [17] to estimate the taxonomical source of each scaffold. Such sources may be helpful to identify putative biological contaminants. Additionally, MOSGA 2 includes NCBI's VecScreen based on the UniVec database to identify adapter sequences.

2.6. Organellar DNA scanner

MOSGA 2 is optimized for annotating nuclear DNA sequences from eukaryotic cells and is less suited for organellar DNA, such as mitochondrial and plastid genomes. To identify organellar DNA in genome assemblies, MOSGA 2 combines information from GC–content, plastid and mitochondrial reference protein databases with RNA prediction tools such as barrnap and tRNAscan-SE 2.0 [6]. At the end of each annotation job, MOSGA 2 creates a relative ranking with the most likely scaffolds. The scoring is an arithmetic calculation based on the density of organellar-specific genes, the numbers of tRNAs, rRNAs, and the GC–content variance. To create the plastid and mitochondria reference proteins databases, we clustered respective RefSeq [23] databases with MMSeq2 [39] and only kept representative sequences of these clusters into our databases to remove redundancy.

2.7. Taxonomy search

In several MOSGA annotation jobs, we observed that multiple users did not select the best suitable gene-prediction models for the given data. Depending on the selection of gene prediction tools, this task could be challenging since, for example, the gene predictor Augustus [37] currently includes already 80 species-specific models. Identifying the most suitable models requires knowledge or an educated guess about each listed species and their relatedness. To support the user during this task, we implemented a taxonomy search for taxonomy-related options. To do so, users select the species name for the uploaded genome assemblies and MOSGA 2 searches for each tool's best putative species- or lineage-specific parameter. Internally, MOSGA 2 contains a trimmed version of the NCBI taxonomy database [33] and searches for the shortest weighted distance between two given nodes. This feature is available for the gene-prediction tools Augustus, GlimmerHMM [20], and SNAP [14] and the validation tool BUSCO.

2.8. Annotation quality

By default, each finished MOSGA genome annotation is validated by NCBI's tbl2asn. In MOSGA 2, we inserted additional multiple filters that improve NCBI compatibility, mainly following Pirovano *et al.* [28]. We integrated additional filters checking the suggested sizes for exons, introns, and the completeness of protein-coding sequences; this includes internal stop-codons and correct start- and stop-codons.

2.9. Integration of existing annotations

MOSGA 2 can import existing genome annotations in GenBank flat format (GBFF) and, therefore, can combine or complete existing annotations with output from additional prediction tools. Results from prediction tools can be visually compared using JBrowse [4]. The GBFF file support is not limited to annotation jobs, but can also be used for comparative genomics tasks or to mix different file formats, which are subsequently interpreted by our tool.

2.10. External application programming interfaces

As another new feature, we introduced three APIs to established external tools: g:Profiler g:GOST [29] for functional enrichment analysis, Integrated Interactions Database [15], and the STRING database [41] for Protein–Protein Interactions Analysis. MOSGA 2 submits predicted protein identifiers from functional annotation to these tools by enabling multiple APIs in the annotation mode and passing the results back into the job submission.

2.11. Gene-prediction

To improve one of the main tasks of MOSGA, we included two new workflows to predict protein-coding genes with BRAKER 2 [3]. New genes can be found based on protein- or orthology-based evidence derived from the OrthoDB database.

3. Results

To demonstrate the usability of MOSGA 2, we will demonstrate several features with exemplary cases.

3.1. Phylogenetics, genome completeness and Average Nucleotide Identity

We performed a phylogenetic analysis based on seven different genome assemblies from the *Saccharomyces* genus (see Table S1). BUSCO served as the source for gene detection with the Eukaryota lineage OrthoDB dataset. MOSGA 2 concatenated the genes, produced a multiple-sequence alignment with MAFFT, and trimmed the MSA with trimAl. The trimmed MSA is used to calculate the phylogenetics with RAxML. The resulting phylogenetic tree in Fig. 1 displays a branching topology that is identical to previous analyses [27]. We marked *S. uvarum* as an outgroup to define the tree rooting. Since BUSCO provided the gene source for the phylogenetic analysis in this example, MOSGA 2 evaluated the genome completeness for each genome that is shown in Fig. 2. The distribution of common missed and unique missed BUSCOs for each genome is shown in Fig. S2. Most of the missing BUSCOs were common to all genomes, indicating that the Eukaryota lineages do not entirely cover all species from the *Saccharomyces* genus or that these orthologs are indeed absent from *Saccharomyces* genomes. According to the BUSCO genome completeness analysis shown in Fig. 2, the genome assembly of *S. mikatae* seems to be incomplete. This could be confirmed by applying an additional EukCC analysis for genome completeness (see Fig. S1). The incompleteness of *S. mikatae* is likely to be derived from the high number of scaffolds and indicates a low-quality genome assembly. In this context, EukCC detected a maximum silent contamination value over 40% for this assembly. Furthermore, MOSGA 2 calculated the Average Nucleotide Identity that is shown in Fig. 2. *S. cerevisiae*'s genome shows a high similarity with to *S. paradoxus* genome and *S. eubayanus* to *S. uvarum*. Indeed, phylogenetic tree analysis indicates a closer relatedness of these species.

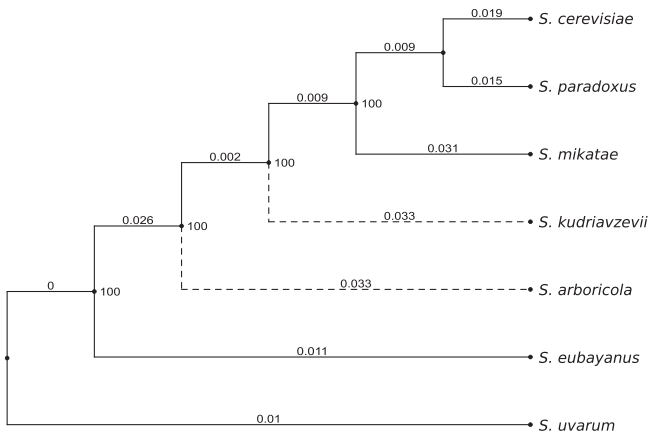

Fig. 1. Multi-gene phylogenetic tree of seven Saccharomyces species. This phylogenetic tree output of MOSGA 2 was computed using BUSCO MAFFT, traimAl, and RAxML based on 173 identified BUSCO genes. The tree topologies are identical to the one described in [27]. Dashed lines indicate a higher distance than th.e 90th quantile.

3.2. Protein-coding genes comparison

To demonstrate the protein-coding gene comparison feature of MOSGA 2, we used Augustus v3.4.0 to predict protein-coding sequences in *S. paradoxus* and five *S. cerevisiae* strains. The chosen assemblies are listed in Table S2. Protein-coding prediction consistency can be checked independently from the used software, especially for annotating different strains from the same species or identifying contaminant species. Depending on the selected threshold for acceptance of a gene match, the strictness increases, and the genes for each genome are mostly only matching with their original genome, as shown in Fig. 3. By decreasing the identity threshold, the total matches became more undifferentiable. Yet, it is important to note that it is possible to differentiate between *S. paradoxus* and *S. cerevisiae* genes similarities. Additionally, we could observe that *S. cerevisiae* strain SK1 has more gene similarities with S288C, while HLJ167, Y12 and sake001 are more similar to each other. We could confirm this observation by calculating the phylogenetic tree with 1873 BUSCOs for these genomes based on the BUSCO *Saccharomycetes* OrthoDB data source, shown in Fig. S3.

3.3. Organellar DNA scanner

We applied our organellar DNA scanner on twenty diverse eukaryotic genomes to quickly identify the mitochondrial, chloroplastid, or other organelles' scaffolds. We chose genome assemblies from various eukaryotic taxa, including plants, nematodes, protists, fungi, vertebrates, and mammalians, to evaluate our scoring. For that purpose, we performed for each genome individual annotations jobs with default settings but without any protein-coding gene prediction. The results of the organellar DNA scanner are represented in a table containing all scaffolds and putative indicators for the presence of an organellar scaffold. To facilitate the interpretation of this table, we introduced a simple scoring system that considers all putative indicators. The highest-ranking score sorts the resulting scoring table rows, shown for *Nannochloropsis oceanica* in Table S3. In 16 out of 20 assemblies, the first hit belonged to one of the organellar scaffolds. An overview of the results is shown in Table 1 and a detailed table considering the type of the organelles is represented in Table S4. The scoring generally performs worse if the number of unplaced contigs increases. We validated the positive matches by comparing our identified first-hit scaffold with the NCBI Genome database.

3.4. Taxonomy search

The integrated taxonomy search in MOSGA 2 tries to identify the most suitable model or lineage for the annotation workflow. Depending on the completeness of the NCBI taxonomy and the available models, it can quickly identify the putative best parameter for each tool. For example, defining *Saccharomyces cerevisiae* as the input genome species, MOSGA 2 selects the *Saccharomyces* Augustus species model and the *Saccharomycetes* OrthoDB data

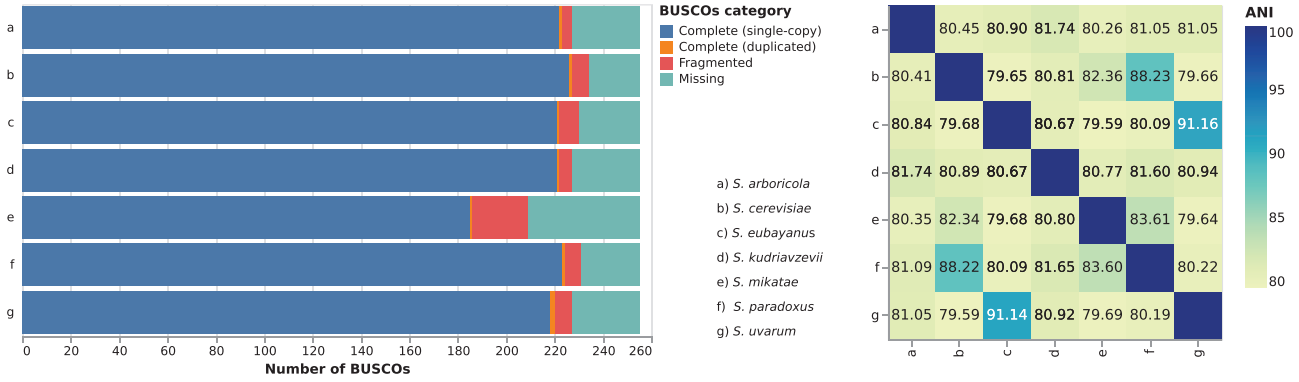

Fig. 2. Merged visualization of BUSCO and FastANI results. Visual representation from the BUSCO analysis based on the Eukaryota OrthoDB lineage shows the genome-completeness of seven different yeast species and the heatmap from Average Nucleotide Identity (ANI) analysis.

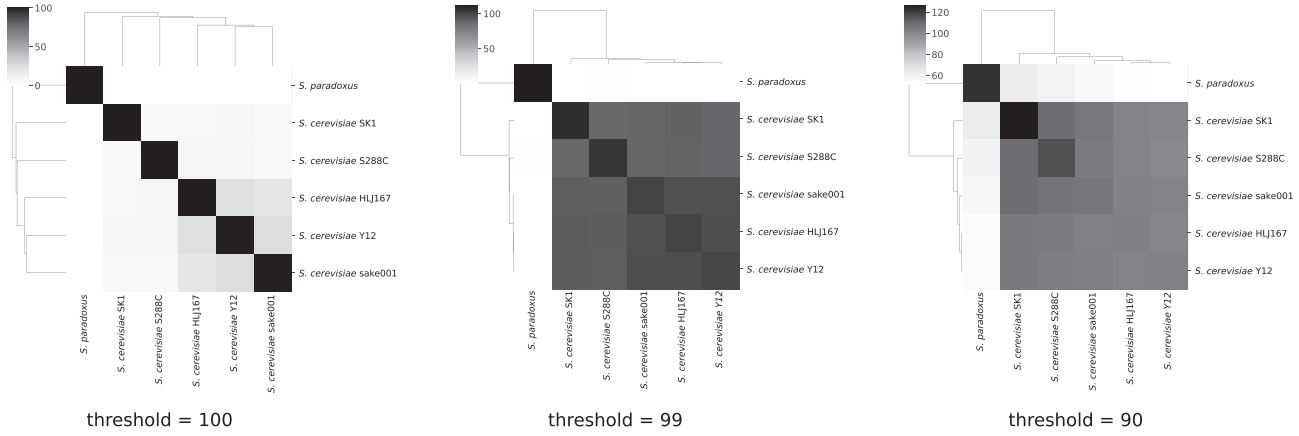

Fig. 3. Gene similarity matrix of six different *Saccharomyces* strains and species based on Augustus predicted protein-coding genes. Depending on the selected threshold for the binning, the similarities differentiate more or less strongly between each genome. Values are relative and normalized by the genome sizes. The heatmap scales are independently defined and differ.

Table 1
Summarized results from the organellar DNA scanner of twenty eukaryotic genome assemblies. **+++** identifies an organellar scaffold as the first suggestion, **++** identifies an organellar scaffold within the first one percent of the suggestions, **+** identifies an organellar scaffold within the first five percent of the suggestions. The total result ranking is shown in Table S3.

Species	No. Scaffolds	Result
Arabidopsis thaliana	7	+++
Apis mellifera	11	+++
Babesia microti	6	+++
Bos taurus	2211	+++
Caenorhabditis elegans	7	+++
Cafeteria burkhardae	170	+++
Cardiosporidium cionae	2204	++
Corvus cornix	113	+++
Danio rerio	1917	+++
Drosophila melanogaster	1870	+++
Homo sapiens	164	+++
Ipomoea triloba	17	+++
Nannochloropsis oceanica	32	+++
Plasmodium falciparum	15	+++
Prunus dulcis	692	++
Saccharomyces cerevisiae	17	+++
Salmo salar	241573	++
Strongylocentrotus purpuratus	871	+++
Tribolium castaneum	2149	++
Zea mays	687	+

source for BUSCO. As another example, a search for *Taphrina alni* results in an Augustus model match for *Pneumocystis jirovecii* (Fig. S4). Both species belong to the subphylum *Taphrinomycotina*. The search depends on the completeness of the NCBI-defined taxonomy. In other cases, if the distance is similar, the selected match will be the first hit.

4. Discussion

In MOSGA 2, we streamline the annotation process by providing validation methods and tools. New functionalities enable user-friendly analyses of genome completeness, contamination, or optimal selected parameters. While improving the annotation process, we additionally implemented comparative genomics workflows and thus extended the scope of MOSGA's capabilities.

We demonstrated that MOSGA 2 generates reliable phylogenetic results using the *Saccharomyces* genus as an example. Depending on the chosen parameters, variances have to be expected, but in this analysis, we only used the default MOSGA 2 settings and selected BUSCO as the gene source.

We presented that the protein-coding genes comparison analysis helps to identify differences in samples based on previously predicted protein genes or inconsistent gene predictions.

Furthermore, besides quality controls such as the genome completeness analysis, we showed that MOSGA 2 could facilitate data preparation by identifying organellar scaffolds inside a genome assembly. In most cases, MOSGA 2 could identify the organellar scaffold directly, although this depends on the assembly quality. The precision of the organellar DNA scanner decreases in samples with many scaffolds, which could indicate problems with biologi-

cal contamination. Moreover, we demonstrated that the taxonomy search feature supports the user in finding the most appropriate model. This search generally relies heavily on NCBI taxonomy quality. This is especially relevant when taxonomical information is incomplete.

During MOSGA 2 development, we implemented feedback and observations from MOSGA users, such as wrongly chosen species models or eukaryotic assemblies with organellar scaffolds. MOSGA 2 improved in terms of quality and user-friendliness by implementing validation tools and functions like the taxonomy search and the organellar DNA scanner. Furthermore, we integrated new workflows that allow comparative genomics analyses and expand the scope of MOSGA analysis for a wider range of applications. Our comparative genomics workflows are based on state-of-the-art tools. For output visualization, we have followed the guidelines of Hattab *et al.* [10], and for improved genome annotation quality and validation tools, the advice of Pirovano *et al.* [28]. In total, we increased the number of implemented workflow rules from 63 to 129. The basic workflows are illustrated in Fig. S5. Figs. S6 and S7 show two job examples for comparative genomics and the annotation workflow. Although MOSGA 2 grows in complexity, it relies on our established architecture that allows high scalability and reproducibility thanks to the multiple-layer design and the Snakemake workflow engine [21]. We implemented several Snakemake and comprehensive data analysis tests that GitLab performs on source-code contribution to maintain the code quality. We encourage scientists to send us implementation and workflow suggestions to support the development of MOSGA 2.

Data availability

Our source code is MIT-licensed freely available on GitLab (gitlab.com/mosga/mosga) and Zenodo (doi: 10.5281/zenodo.5121228). An online version of MOSGA 2 is available under mosga.mathematik.uni-marburg.de. This website is free and open to all users, and there are no registration requirements. The calculated results for the given examples are available under mosga.mathematik.uni-marburg.de/phylo for the *Saccharomyces* phylogenetics example, mosga.mathematik.uni-marburg.de/genecomp for the *Saccharomyces* strains gene comparison and mosga.mathematik.uni-marburg.de/organelles for the organellar DNA scanner.

Funding

This work was supported by the LOEWE program of the State of Hesse (Germany) in the MOSLA research cluster.

Author contributions statement

R.M. wrote the manuscript, designed and developed the framework. H.D. implemented the comparative genomics workflows and assisted in the development. G.H. supported the tool development and data visualizations, he proofread and revised the manuscript. T.H. provided the database for the organellar DNA scanner and ideas for implementation and revised the manuscript. M.G.F. revised the manuscript. D.H. supervised the project, discussed the results, and revised the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2021.09.024.

References

[1] Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, et al. Sci Data 2018;5(1):180154. https://doi.org/10.1038/sdata.2018.154. ISSN 2052-4463. url: http://www.nature.com/articles/sdata2018154.
[2] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Nat Biotechnol 2017;35(8):725–31. https://doi.org/10.1038/nbt.3893. ISSN 1087-0156. url: http://www.nature.com/articles/nbt.3893.
[3] Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform 2021;3(1):1–11. https://doi.org/10.1093/nargab/lqaa108. URL: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa108/6066535.
[4] Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. Genome Biol 2016;17(1):1–12. https://doi.org/10.1186/s13059-016-0924-1. ISSN 1474760X.
[5] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348. url: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp348.
[6] Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Research 2021;49(16):9077–96. https://doi.org/10.1093/nar/gkab688. url: https://academic.oup.com/nar/article/49/16/9077/6355886.
[7] del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol 2014;29(5):252–9. https://doi.org/10.1016/j.tree.2014.03.006. url: https://linkinghub.elsevier.com/retrieve/pii/S0169534714000640.
[8] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 2004;32(5):1792–7. https://doi.org/10.1093/nar/gkh340. ISSN 1362-4962.
[9] Hackl T, Martin R, Barenhoff K, Duponchel S, Heider D, Fischer MG. Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria roenbergensis. Sci Data 2020;7(1). https://doi.org/10.1038/s41597-020-0363-4. url: http://www.nature.com/articles/s41597-020-0363-4.
[10] Hattab G, Rhyne T-M, Heider D. Ten simple rules to colorize biological data visualization. PLOS Comput Biol 2020;16(10):e1008259. https://doi.org/10.1371/journal.pcbi.1008259. ISSN 1553-7358.
[11] Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9(1):1–8. https://doi.org/10.1038/s41467-018-07641-9. ISSN 20411723.
[12] Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res 2002;30(14):3059–66. https://doi.org/10.1093/nar/gkf436. ISSN 13624962. url: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf436.
[13] Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010.
[14] Korf I. Gene finding in novel genomes. BMC Bioinform 5 (2004) 59. ISSN 14712105. doi:10.1186/1471-2105-5-59. url:http://www.ncbi.nlm.nih.gov/pubmed/15144565 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC421630.
[15] Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species, Nucleic Acids Research 47(D1): D581–D589, Jan 2019. ISSN 0305-1048. doi:10.1093/nar/gky1037. url: https://academic.oup.com/nar/article/47/D1/D581/5165345.
[16] Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucl Acids Res 2019;47(D1):D807–11. https://doi.org/10.1093/nar/gky1053. url:http://www.ncbi.nlm.nih.gov/pubmed/30395283 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6323947.
[17] Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. F1000Research 2017;6:1287. https://doi.org/10.12688/f1000research.12232.1. ISSN 2046-1402. url: https://f1000research.com/articles/6-1287/v1.

[18] Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, et al. Bioinformatics 2007;23(21):2947–8. https://doi.org/10.1093/bioinformatics/btm404. ISSN 1367-4803. url: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm404.
[19] Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evolut 2015;32 (10):2798–800. https://doi.org/10.1093/molbev/msv150. url: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv150 http://www.ncbi.nlm.nih.gov/pubmed/26130081 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4576710.
[20] Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics 2004;20(16):2878–9. https://doi.org/10.1093/bioinformatics/bth315. ISSN 13674803.
[21] Martin R, Hackl T, Hattab G, Fischer MG, Heider D. MOSGA: modular open-source genome annotator. Bioinformatics 2021;36(22–23):5514–5. https://doi.org/10.1093/bioinformatics/btaa1003. ISSN 1367-4803. url: http://www.ncbi.nlm.nih.gov/pubmed/33258916 https://academic.oup.com/bioinformatics/article/36/22-23/5514/6015104.
[22] Mi H, Muruganujan A, Thomas PD, function Modeling the evolution of gene, attributes other gene. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucl Acids Res 2013;41(Database issue):D377–86. https://doi.org/10.1093/nar/gks1118. url:http://www.ncbi.nlm.nih.gov/pubmed/23193289 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531194.
[23] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucl Acids Res 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189. ISSN 13624962.
[24] Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ et al. Charting the complexity of the marine microbiome through single-cell genomics. Cell, 179 (7): 1623–1635.e11; 2019. ISSN 1097–4172. doi:10.1016/j.cell.2019.11.017. url:http://www.ncbi.nlm.nih.gov/pubmed/31835036 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6919566.
[25] Palfalvi G, Hackl T, Terhoeven N, Shibata TF, Nishiyama T, Ankenbrand M, et al. Curr Biol 2020;30(12). https://doi.org/10.1016/j.cub.2020.04.051. 2312–2320.e5, ISSN 09609822.https://doi.org/10.1016/j.cub.2020.04.051. url:https://linkinghub.elsevier.com/retrieve/pii/S0960982220305674.
[26] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics (Oxford, England), 35(3);526–528:2019. ISSN 1367-4811. doi:10.1093/bioinformatics/bty633.url:http://www.ncbi.nlm.nih.gov/pubmed/30016406.
[27] Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature 2018;556 (7701):339–44. https://doi.org/10.1038/s41586-018-0030-5. ISSN 1476-4687. url:http://www.ncbi.nlm.nih.gov/pubmed/29643504 ttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6784862.
[28] Pirovano W, Boetzer M, Derks MF, Smit S. NCBI-compliant genome submissions: tips and tricks to save time and money. Briefings Bioinform 2017;18(2):179–82. https://doi.org/10.1093/bib/bbv104. ISSN 14774054.
[29] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucl Acids Resarch 2019;47(W1):W191–8. https://doi.org/10.1093/nar/gkz369. ISSN 0305-1048. url:https://academic.oup.com/nar/article/47/W1/W191/5486750.
[30] Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. Genome Biol 2020;21 (1):244. https://doi.org/10.1186/s13059-020-02155-4. ISSN 1474-760X. url: http://www.ncbi.nlm.nih.gov/pubmed/32912302 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7488429 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02155-4.
[31] Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res 2012;22(3):557–67. ISSN 1549-5469. url:http://www.ncbi.nlm.nih.gov/pubmed/22147368 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3290791.
[32] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucl Acids Res 2019;47(D1):D94–9. https://doi.org/10.1093/nar/gky989. ISSN 13624962.
[33] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database 2020;2020(2):1–21. https://doi.org/10.1093/database/baaa062. ISSN 17580463.
[34] Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, et al. Hidden diversity of soil giant viruses. Nat Commun 2018;9(1):4881. https://doi.org/10.1038/s41467-018-07335-2. ISSN 2041-1723. url: http://www.nature.com/articles/s41467-018-07335-2.
[35] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31(19):3210–2. ISSN 1367-4803. url:https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv351.
[36] .Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England), 30 (9): 1312–3, May 2014. ISSN 1367–4811. doi:10.1093/bioinformatics/btu033. url:https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu033 http://www.ncbi.nlm.nih.gov/pubmed/24451623 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3998144.
[37] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucl Acids Res 33 (Web Server issue): W465–7, Jul 2005. ISSN 1362–4962. doi:10.1093/nar/gki458. url:http://www.ncbi.nlm.nih.gov/pubmed/15980513 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1160219.
[38] Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biology 2020;18(12):e3001007. ISSN 1545-7885. doi:10.1371/journal.pbio.3001007. url:https://doi.org/10.1371/journal.pbio.3001007 https://dx.plos.org/10.1371/journal.pbio.3001007.
[39] Steinegger M, Söding J. MMsEqs 2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35 (11):1026–8. https://doi.org/10.1038/nbt.3988. ISSN 15461696.
[40] Sun L, Gao T, Wang F, Qin Z, Yan L, Tao W et al. Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology. Mol Ecol Resour, pages 1755–0998.13190, Jul 2020. ISSN 1755–098X. doi:10.1111/1755-0998.13190. url:https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13190.
[41] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucl Acids Res 2019;47(D1):D607–13. ISSN 0305-1048. doi:10.1093/nar/gky1131. url:https://academic.oup.com/nar/article/47/D1/D607/5198476.
[42] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 1994;22(22):4673–80. https://doi.org/10.1093/nar/22.22.4673. ISSN 0305-1048. url:https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/22.22.4673.
[43] Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 2018;35(3):543–8. https://doi.org/10.1093/molbev/msx319. ISSN 15371719. doi:10.1093/molbev/msx319.
[44] Wiley G, Miller MJ. A Highly Contiguous Genome for the Golden-Fronted Woodpecker (Melanerpes aurifrons) via Hybrid Oxford Nanopore and Short Read Assembly. G3: Genes—Genomes—Genetics, 10 (6): 1829–1836, jun 2020. ISSN 2160–1836. doi:10.1534/g3.120.401059. url:http://g3journal.org/lookup/doi/10.1534/g3.120.401059.
[45] Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evolut, 8(1): 28–36, Jan 2017. ISSN 2041–210X. doi:10.1111/2041-210X.12628. url:https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628.
[46] Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 25(5) 656–667.e8, May 2019. ISSN 19313128. doi:10.1016/j.chom.2019.03.007. url:https://linkinghub.elsevier.com/retrieve/pii/S1931312819301593.