W132–W137 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 30 April 2022*
https://doi.org/10.1093/nar/gkac298

BusyBee Web: towards comprehensive and differential composition-based metagenomic binning

Georges P. Schmartz [1], Pascal Hirsch[1,2], Jérémy Amand[1,2], Jan Dastbaz[3,4], Tobias Fehlmann [1], Fabian Kern [1,2], Rolf Müller[3,4] and Andreas Keller [1,2,*]

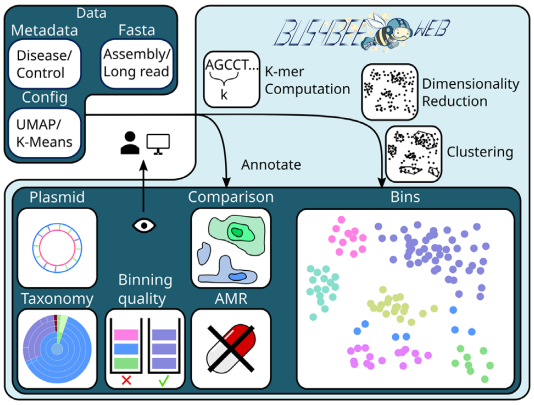
[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Clinical Bioinformatics (CLIB), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany, [3]Microbial Natural Products (MINS), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany and [4]Deutsches Zentrum für Infektionsforschung (DZIF), Standort Hannover-Braunschweig, 38124 Braunschweig, Germany

Received March 08, 2022; Revised April 07, 2022; Editorial Decision April 12, 2022; Accepted April 14, 2022

ABSTRACT

Despite recent methodology and reference database improvements for taxonomic profiling tools, metagenomic assembly and genomic binning remain important pillars of metagenomic analysis workflows. In case reference information is lacking, genomic binning is considered to be a state-of-the-art method in mixed culture metagenomic data analysis. In this light, our previously published tool BusyBee Web implements a composition-based binning method efficient enough to function as a rapid online utility. Handling assembled contigs and long nanopore generated reads alike, the webserver provides a wide range of supplementary annotations and visualizations. Half a decade after the initial publication, we revisited existing functionality, added comprehensive visualizations, and increased the number of data analysis customization options for further experimentation. The webserver now allows for visualization-supported differential analysis of samples, which is computationally expensive and typically only performed in coverage-based binning methods. Further, users may now optionally check their uploaded samples for plasmid sequences using PLSDB as a reference database. Lastly, a new application programming interface with a supporting python package was implemented, to allow power users fully automated access to the resource and integration into existing workflows. The webserver is freely available under: https://www.ccb.uni-saarland.de/busybee.

GRAPHICAL ABSTRACT



INTRODUCTION

State-of-the-art metagenomics data analysis predominantly depends on reference databases. Reads are compared against well-characterized sequences and in case of sufficient sequence similarity, a read may be assigned to a taxonomy, an associated operational taxonomic unit count is incremented, or a genomic function is deduced (1–4). However, metagenomic studies operating at the boundary of what is known to humankind, e.g. investigating extreme maritime or volcanic environments, will inevitably come to the point where reference data is incomplete or of insufficient quality (5–7). While the overall possibilities for analysis are limited, a lack of reference information does not necessarily prevent any analysis. Instead, metagenomic short-read assembly or long-read metagenomic sequencing is frequently performed to allow for further hypothesizing, analysis, and discovery. However, due to high species diversity, sequencing errors, and other conflicts during assembly, metagenomic assemblies frequently yield multiple thousands of contigs of variable lengths and qualities (8,9).

*To whom correspondence should be addressed. Tel: +49 681 30268611; Fax: +49 681 30268610; Email: andreas.keller@ccb.uni-saarland.de

Since short-read metagenomic read assembly and long-read metagenome sequencing output a mix of sequences of all the present species, structured analysis of results remains difficult. Therefore, longer sequences are usually grouped using binning methods to separate sequences into taxonomic units. Two features are frequently used to achieve informed separation into groups. Coverage-based binning uses coverage profiles of sequences, computed across multiple samples, to cluster into bins. Composition-based binning utilizes the conservation of sequence features like tetranucleotide profiles and derives bins from the input sequence (10). Many of the state-of-the-art binning methods such as MaxBin2 are hybrid methods using both kinds of features (11–13). However, coverage profiles provide limited information if only one individual sample is analyzed, and they may even be not applicable depending on the selected sequencing method. Accordingly, new methods that do not require coverage profiles are further developed (14,15).

In 2018, we proposed BusyBee Web as a reference-free composition-based binning tool efficient enough to function as a webserver (16,17). The underlying pipeline trains a classifier on a subset of the input data which is then used to assign sequences into bins. The used features are normalized k-mer profiles of length four or five. The tool optionally provides various functional and taxonomic annotations with Prokka and Kraken respectively allowing for taxonomic binning (18,19). Five years after initial publication, the community used BusyBee Web to analyze >2500 individual samples and perform >4500 runs. Here, we present a major update to the binning resource.

MATERIALS AND METHODS

Developing an update to an existing resource allowed us to revisit some of the already available functionality and cover a broad list of minor improvements. Accordingly, the taxonomic annotation was updated to support Kraken 2 with a newer database and marker genes for bin quality assessment were extended to include the Archaea genes from the anvi'o project (20). Further, a sunburst plot was added and several new expert settings for clustering and embedding methods were implemented. Namely, we included t-SNE (21), FitSNE (22), UMAP (arXiv:1802.03426), PHATE (23) and TriMap (arXiv:1910.00204) as embedding and DBSCAN (24), HDBSCAN (25), *k*-means and spectral clustering (26) as new clustering methods. From the list of new features, we want to highlight three major changes with higher visibility to newer users.

Plasmids annotations

Due to the random sampling involved in shotgun sequencing experiments, metagenomic data often includes plasmid fragments that may also end up in assemblies, potentially impacting downstream analysis. BusyBee Web now optionally compares input sequences to the most recent version of PLSDB using mash screen (27,28). In case plasmid signatures are found, the most relevant information about the plasmids is displayed. From here, users can take a deeper look into the findings by continuing their analysis on PLSDB.

Comparative metagenomics

Group comparison is a frequently requested analysis that is often neglected in composition-based methods. In BusyBee Web, we compute a differential density between two user-defined classes, by first applying a Gaussian 2D kernel to the embedded sequences for both classes separately. Bandwidth and grid size used in the computation can be modified by the user, within given boundaries. Next, the difference between both densities is visualized. This usually results in a picture where various areas are dominated by different classes. While this method does not directly provide statistics on coverage differences, it remains indicative of different phenomena. On the one hand, if long reads are directly embedded, higher density regions should represent a higher relative number of sequences with a similar k-mer spectrum in the sample. On the other hand, if assembled contigs were provided, interpretation becomes more complex. First, the number of embedded sequences is expected to increase simply due to technological errors, resulting in higher density regions for higher sequence counts similar to the long-read interpretation. Second, increased phylogenetic diversity is captured since identical sequences should ideally be collapsed already during assembly. The difference in density can be retrieved for each cluster allowing the user to further analyze potentially interesting patterns and areas.

Application programming interface

To allow programmatic access to BusyBee Web, we implemented an application programming interface (API). The API complies with the Open API 3.0.2 standard (29). Users can start jobs, check their status, and download individual results over the API. Additionally, a python package is supported and distributed via conda, which allows for easy integration into R scripts using reticulate. The package is available on: https://github.com/CCB-SB/busybee_api.

Case studies

In order to benchmark BusyBee Web on a mock community in the first case study, we downloaded the *ERR3152364* dataset from the sequence read archive and converted the fastq files into fasta files while also adapting the header names. Due to the high sequencing depth of the experiment, the sample had to be pruned to comply with the constraints imposed by the webserver. Thus, we shuffled the fasta file randomly and selected the first 200 Mb of data, corresponding exactly to the upload limit and which accounts for <2% of the initial file. The resulting file contained a total of 50 679 reads. After data generation, we started analysis with default parameters changing only the embedding to UMAP. For comparison, various embeddings with different dimension reduction methods were computed (Supplementary Figure S1).

The second case study discussing differences between sequencing technologies was conducted with newly generated data. Both datasets were derived from the same 1mL of bile sample of a healthy human individual and DNA was extracted with the same QiAamp DNA Microbiome Kit allowing for comparison between technolo-

gies. Next-generation sequencing DNA libraries were prepared using the MGIEasy Universal DNA Library Prep set following the recommendations of the manufacturer. The DNBSEQ-G400 was used as short-read sequencing platform. Oxford nanopore sequencing was prepared with the SQK-LSK109 Ligation Sequencing kit before sequencing on an FLO-MIN106D flow cell in a MinION Mk1B. Basecalling was performed with Guppy v5.0.7. For both datasets, human-read contamination was removed by first running kneaddata v0.7.4, followed by sra-human-scrubber v1.0.2021_05_05 (1,30). After removal of human reads, the ONT fastq was converted to fasta and read names were shortened to generic header names. For the short-read sequencing data, reads were assembled to scaffolds with metaSPAdes v3.15.2 and scaffolds were retained (8). Before analysis with BusyBee Web, both datasets, short- and long-read, were combined and a mapping to the original fasta entries was generated. Next, data was passed to BusyBee Web with default settings, but selecting UMAP as embedding algorithm.

RESULTS

With the increasing popularity of whole shotgun metagenome and long-read sequencing competing with amplicon sequencing, dedicated analysis of plasmids from metagenomics data is becoming increasingly tempting to the metagenomics community (31). However, shared sequences between chromosomes and plasmids, variable sizes, and a wide range of other factors render plasmid assembly from short reads an algorithmic challenging task often entailing high misassembly rates (31,32). Similarly, the prediction of both plasmid reads, and plasmid sequences remains an intensively debated field of research, also affecting long-read sequencing technology (33–38). Attributed to these difficulties, plasmid sequences frequently appear in binning inputs where they may be difficult to interpret. With the newly added plasmid annotation, BusyBee Web explicitly notifies the user about the presence of already known putative plasmid signatures. Further, the newly adopted differential density-based visualization allows for visual interpretation of similarity between aggregated samples. Since cohort and interventional studies comparing healthy against diseased patients, elderly against young, or different treatment conditions are increasingly performed in biomedical research, the field also faces an upsurge of comparative metagenomic studies. However, many of the conclusions drawn from cohort studies are either based on differential taxonomic counts or the functional aspect of sequences. In both cases, the comparison relies on reference information. One method to alleviate this constraint is to assess differential coverage profiles of binned sequences. However, similar to coverage-based binning, coverage profiles are required for this approach, which may not be available. Moreover, minor differences in binning outcomes may largely impact conclusions weakening the stability of this approach. The embedding followed by subsequent kernel application that we implemented alleviates these drawbacks and the volatility of results is bound to the characteristics of the selected dimensionality reduction method. Lastly, with the added application programming interface (API) BusyBee Web can easily be integrated into new and existing data analysis pipelines. In combination with workflow managing tools such as Nextflow or Snakemake, the API increases experiment throughput and reproducibility of results (39,40).

In order to highlight the improved functionality of BusyBee Web, we analyzed two datasets of varying ground truth information. While at the core BusyBee uses a reference-free algorithm for binning, here we make use of reference-based taxonomic annotations that were added after binning. Combining these annotations with the knowledge of well-characterized microbial environments allows us to better gauge binning quality.

Mock community benchmark

To assess the binning quality of BusyBee Web on a well-characterized example, we used a dataset by Nicholls et al. as ground truth (41). This nanopore sequencing data represents a mock community composed of exactly ten known species. The output of BusyBee Web consists of 27 bins (Figure 1A and B). However, 14 of these bins each contained <1% of sequences and may be discarded from further analysis. Of the remaining 13 bins, five bins, namely 1, 5, 8, 22 and 24, were mostly composed of unclassified sequences. We postulate that these bins are mostly made up of *Cryptococcus neoformans* and *Saccharomyces cerevisiae* which are not included in the selected Kraken 2 database. The taxonomic composition of bin 9, which is the smallest remaining bin composing only 515 sequences, is highly fractioned indicating a low binning quality. While not exempt from cross-contamination, all the remaining bins (2, 11, 13, 16, 17, 21 and 27) can clearly be attributed to the distinct species from the mock community, indicating that despite only using a fraction of the input, BusyBee Web is able to successfully recover the contained major species.

Sequencing technology comparison

To highlight the new analysis functionality added in this update, we compared the suitability of long reads with short-read assembled scaffolds. With the 19,262 input sequences passing the default length filter a total of 19 bins were predicted of which five (5, 7, 8, 15 and 19) contained <1% of sequences. A total of 340 sequence similarities to potentially relevant plasmids were identified where the majority was reported in *Enterococcus faecium*. Looking at the new differential density plot from Figure 1C we observe six clusters (1, 2, 4, 6, 16 and 17) that are specific to the short-read sequencing experiment. The taxonomic profile of cluster 1 has a high relative number of unclassified sequences, pointing towards potentially unreliable assemblies (Figure 1D). Nevertheless, we note that within this bin a few long reads were found at a relative proportion of ∼15.5%. Four of the remaining five clusters (2, 4, 16 and 17) have low contaminations. These four clusters presumably consist mostly of Lachnospiraceae, *Enterobacteriaceae*, *Actinomycetaceae* and *Micrococcaceae* respectively. Potentially, due to biological random sampling or decreased sequencing depth, these genomic signatures mostly escaped the nanopore sequencing.

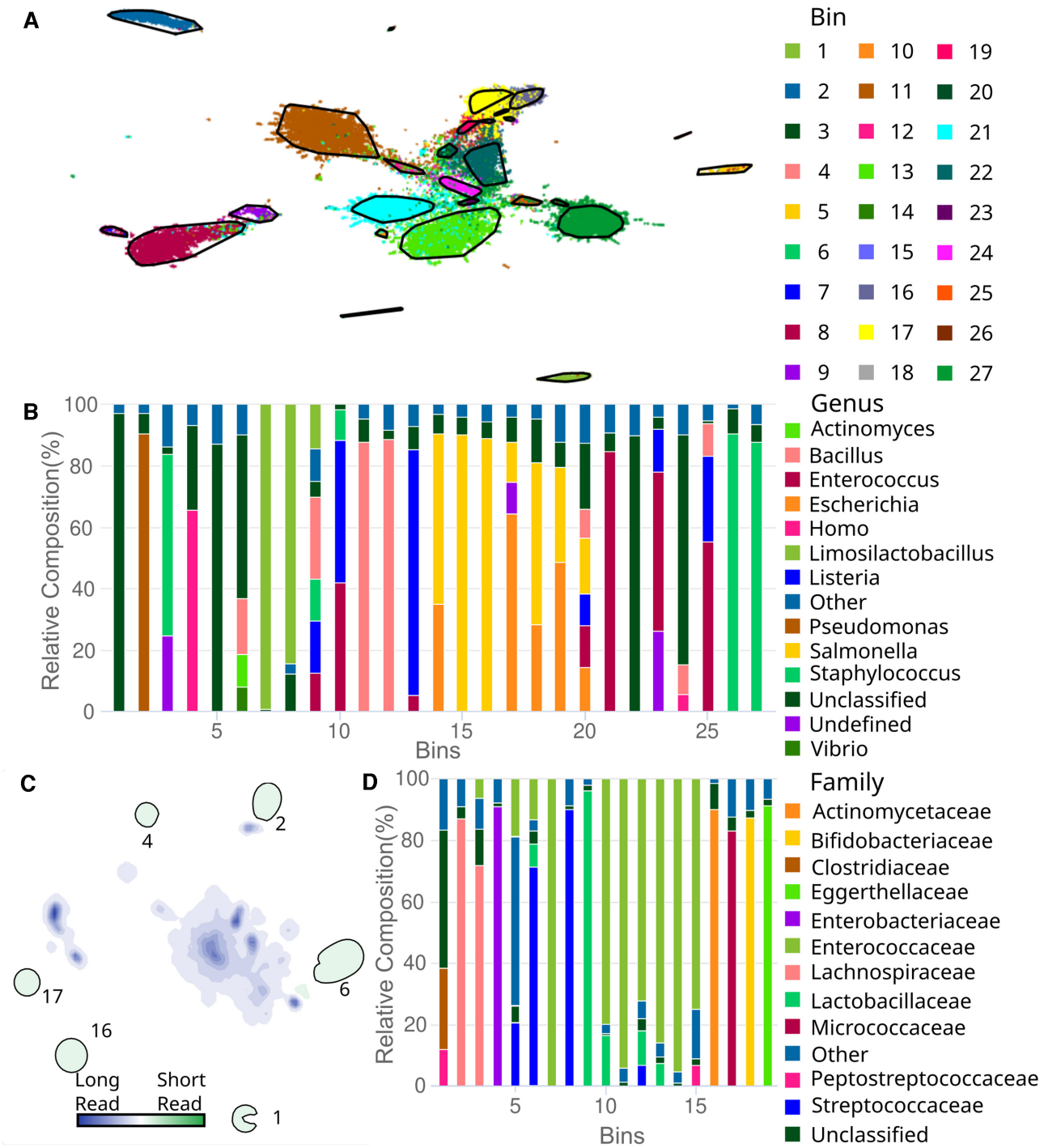


Figure 1. (**A**) Embedding of the mock community dataset, using UMAP with default settings. (**B**) Taxonomic profile at genus level of the different bins computed on a mock community composed of ten different species. (**C**) Differential density embedding of a bile sample sequenced with Oxford Nanopore MinION (Long Read) and DNBSEQ-400 (Short Read) respectively. (**D**) Taxonomic annotation of bins computed on the comparison dataset.

CONCLUSION

With the new update, we substantially extended the capabilities of BusyBee Web as a versatile composition-based binning tool. On the one hand, with the newly added clustering methods, embedding algorithms, and API, we increased the data analysis possibilities for expert users. On the other hand, we hope to widen our user base by providing new visualizations and annotations. While we always strive for maximal flexibility, the ease of use of BusyBee Web as an installation-free webservice comes at a cost. For example, the data upload is limited to 200Mb per sample which can quickly be reached if multiple samples are being analyzed. Moreover, some of the presented clustering and embedding options will not be able to handle the theoret-

ical maximal number of contigs that fit into a 200Mb file, due to time and memory constraints. Therefore, BusyBee Web provides an option for compressing information before embedding computation, alleviating some of these limitations. Nonetheless, visualization of the embedding in the local browser for many data points may become slow or irresponsive on less powerful hardware. Here, we recommend to prefer API usage instead. Moreover, with sufficient coverage information available, state-of-the-art coverage-based and hybrid metagenomic binning tools are expected to outperform composition-based tools on short-read sequencing data in larger projects.

Potential future development efforts may further focus on the identification of mobile genetic elements. However, with large disagreements already observed across plasmid classification tools, potential counter-strategies, e.g. automated removal of putative sequences from user input, are likely unstable and thus currently not advisable. Further, by extending the BusyBee Web server to allow for a selection of different embedding and clustering methods, it will be easier in the future to integrate newer algorithms into the generalized framework.

DATA AVAILABILITY

BusyBee Web is freely available at: https://www.ccb.uni-saarland.de/busybee.

ACCESSION NUMBERS

Respecting the German federal privacy law, we uploaded the short- and long-read data after human read removal to the Sequence Read Archive. Preprocessed data can be found in NCBI SRA using the accession numbers SRX14022915 and SRX14435297.

The mock community dataset was made available by Nicholls et al. in the Sequence Read Archive under the accession: ERR3152364.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the UdS-HIPS TANDEM Initiative. We thank Valentina Galata for fruitful discussions on potential case studies.
G.P.S.: Software, Formal Analysis, Writing - Original Draft, Visualization; P.H.: Software, Writing - Review & Editing; J.A.: Methodology, Software, Writing - Review & Editing; J.D.: Investigation, Data Curation; F.K.: Writing - Review & Editing, Supervision; T.F.: Software, Supervision; R.M.: Conceptualization, Resources, Funding acquisition; A.K.: Conceptualization, Resources, Supervision, Project administration, Funding acquisition.

FUNDING

Funding for open access charge: Saarland University.
Conflict of interest statement. G.P.S., R.M., and A. K. are shareholders of MOOH GmbH.

REFERENCES

1. Beghini,F., McIver,L.J., Blanco-Miguez,A., Dubois,L., Asnicar,F., Maharjan,S., Mailyan,A., Manghi,P., Scholz,M., Thomas,A.M. *et al.* (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, **10**, e65088.
2. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with kraken 2. *Genome Biol.*, **20**, 257.
3. Bharti,R. and Grimm,D.G. (2021) Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform*, **22**, 178–193.
4. Milanese,A., Mende,D.R., Paoli,L., Salazar,G., Ruscheweyh,H.J., Cuenca,M., Hingamp,P., Alves,R., Costea,P.I., Coelho,L.P. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1014.
5. Spieck,E., Spohn,M., Wendt,K., Bock,E., Shively,J., Frank,J., Indenbirken,D., Alawi,M., Lucker,S. and Hupeden,J. (2020) Extremophilic nitrite-oxidizing chloroflexi from yellowstone hot springs. *ISME J.*, **14**, 364–379.
6. Wibowo,M.C., Yang,Z., Borry,M., Hubner,A., Huang,K.D., Tierney,B.T., Zimmerman,S., Barajas-Olmos,F., Contreras-Cubas,C., Garcia-Ortiz,H. *et al.* (2021) Reconstruction of ancient microbial genomes from the human gut. *Nature*, **594**, 234–239.
7. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.
8. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
9. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, **31**, 1674–1676.
10. Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glockner,F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.
11. Kang,D.D., Li,F., Kirton,E., Thomas,A., Egan,R., An,H. and Wang,Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.
12. Mallawaarachchi,V., Wickramarachchi,A. and Lin,Y. (2020) GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, **36**, 3307–3313.
13. Wu,Y.W., Simmons,B.A. and Singer,S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.
14. Wickramarachchi,A. and Lin,Y. (2021) *21st International Workshop on Algorithms in Bioinformatics (WABI 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
15. Wickramarachchi,A., Mallawaarachchi,V., Rajan,V. and Lin,Y. (2020) MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics*, **36**, i3–i11.
16. Laczny,C.C., Kiefer,C., Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2017) BusyBee web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.*, **45**, W171–W179.
17. Benson,G. (2017) Editorial: the 15th annual nucleic acids research web server issue 2017. *Nucleic Acids Res.*, **45**, W1–W5.
18. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
19. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
20. Eren,A.M., Esen,O.C., Quince,C., Vineis,J.H., Morrison,H.G., Sogin,M.L. and Delmont,T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319.
21. Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
22. Linderman,G.C., Rachh,M., Hoskins,J.G., Steinerberger,S. and Kluger,Y. (2019) Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods*, **16**, 243–245.
23. Moon,K.R., van Dijk,D., Wang,Z., Gigante,S., Burkhardt,D.B., Chen,W.S., Yim,K., Elzen,A.V.D., Hirn,M.J., Coifman,R.R. *et al.*

(2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, **37**, 1482–1492.
24. Xu,X., Ester,M., Kriegel,H.-P. and Sander,J. (1998), *Proceedings 14th International Conference on Data Engineering*. IEEE, pp. 324–331.
25. Campello,R.J., Moulavi,D. and Sander,J. (2013), *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 160–172.
26. Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
27. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.*, **17**, 132.
28. Schmartz,G.P., Hartung,A., Hirsch,P., Kern,F., Fehlmann,T., Muller,R. and Keller,A. (2022) PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.*, **50**, D273–D278.
29. Tarkowska,A., Carvalho-Silva,D., Cook,C.E., Turner,E., Finn,R.D. and Yates,A.D. (2018) Eleven quick tips to build a usable REST API for life sciences. *PLoS Comput. Biol.*, **14**, e1006542.
30. Katz,K.S., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2021) STAT: a fast, scalable, minhash-based k-mer tool to assess sequence read archive next-generation sequence submissions. *Genome Biol.*, **22**, 270.
31. Antipov,D., Raiko,M., Lapidus,A. and Pevzner,P.A. (2019) Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.*, **29**, 961–968.
32. Pellow,D., Zorea,A., Probst,M., Furman,O., Segal,A., Mizrahi,I. and Shamir,R. (2021) SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*, **9**, 144.
33. Krawczyk,P.S., Lipinski,L. and Dziembowski,A. (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.
34. Laczny,C.C., Galata,V., Plum,A., Posch,A.E. and Keller,A. (2019) Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief Bioinform*, **20**, 857–865.
35. Pellow,D., Mizrahi,I. and Shamir,R. (2020) PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.*, **16**, e1007781.
36. Pradier,L., Tissot,T., Fiston-Lavier,A.S. and Bedhomme,S. (2021) PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinf.*, **22**, 349.
37. Wickramarachchi,A. and Lin,Y. (2022) GraphPlas: refined classification of plasmid sequences using assembly graphs. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **19**, 57–67.
38. Zhou,F. and Xu,Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, **26**, 2051–2052.
39. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
40. Molder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A. *et al.* (2021) Sustainable data analysis with snakemake. *F1000Res.*, **10**, 33.
41. Odahara,M., Nakamura,K., Sekine,Y. and Oshima,T. (2021) Ultra-deep sequencing reveals dramatic alteration of organellar genomes in physcomitrella patens due to biased asymmetric recombination. *Commun. Biol.*, **4**, 633.