Community-led, integrated, reproducible multi-omics with anvi'o

A. Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, Jessica N. Pan, Mahmoud Yousef, Emily C. Fogarty, Florian Trigodet, Andrea R. Watson, Özcan C. Esen, Ryan M. Moore, Quentin Clayssen, Michael D. Lee, Veronika Kivenson, Elaina D. Graham, Bryan D. Merrill, Antti Karkman, Daniel Blankenberg, John M. Eppley, Andreas Sjödin, Jarrod J. Scott, Xabier Vázquez-Campos, Luke J. McKay, Elizabeth A. McDaniel, Sarah L. R. Stevens, Rika E. Anderson, Jessika Fuessel, Antonio Fernandez-Guerra, Lois Maignien, Tom O. Delmont, Amy D. Willis

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., ... & Willis, A. D. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. Nature microbiology, 6(1), 3-6.3

Community-led, integrated, reproducible multi-omics with anvi'o

A. Murat Eren[1,2,3,*], Evan Kiefl[1,4], Alon Shaiber[1,4], Iva Veseli[1,4], Samuel E. Miller[1], Matthew S. Schechter[1,2], Isaac Fink[1], Jessica N. Pan[1], Mahmoud Yousef[1], Emily C. Fogarty[1], Florian Trigodet[1], Andrea R. Watson[1], Özcan C. Esen[1], Ryan M. Moore[5], Quentin Clayssen[6], Michael D. Lee[7,8], Veronika Kivenson[9], Elaina D. Graham[10], Bryan D. Merrill[11], Antti Karkman[12], Daniel Blankenberg[13,14], John M. Eppley[15], Andreas Sjödin[16], Jarrod J. Scott[17], Xabier Vázquez-Campos[18], Luke J. McKay[19,20], Elizabeth A. McDaniel[21], Sarah L. R. Stevens[22,23], Rika Anderson[24], Jessika Fuessel[1], Antonio Fernandez-Guerra[25], Lois Maignien[3,26], Tom O. Delmont[27], Amy D. Willis[28]

[1] Department of Medicine, University of Chicago, Chicago, IL, USA; [2] Committee on Microbiology, University of Chicago, Chicago, IL, USA; [3] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA; [4] Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL, USA; [5] Center for Bioinformatics and Computational Biology, University of Delaware, DE, USA; [6] Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland; [7] Exobiology Branch, NASA Ames Research Center, Mountain View, CA, USA; [8] Blue Marble Space Institute of Science, Seattle, WA, USA; [9] Department of Microbiology, Oregon State University, Corvallis, OR, USA; [10] Department of Biological Sciences, University of Southern California, CA, USA; [11] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA; [12] Department of Microbiology, University of Helsinki, Helsinki, Finland; [13] Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA; [14] Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA; [15] Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Manoa, Honolulu, HI, USA; [16] Division of CBRN Security and Defence, Swedish Defence Research Agency - FOI, Umeå, Sweden; [17] Smithsonian Tropical Research Institute, Bocas del Toro, Republic of Panamá; [18] NSW Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, NSW 2052, Australia; [19] Center for Biofilm Engineering, Montana State University, Bozeman, MT, USA; [20] Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT, USA; [21] Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; [22] Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA; [23] American Family Insurance Data Science Institute, University of Wisconsin-Madison, Madison, WI, USA; [24] Department of Biology, Carleton College, Northfield, MN, USA; [25] Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark; [26] Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Univ Brest, CNRS, Ifremer, Plouzané, France; [27] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France; [28] Department of Biostatistics, University of Washington, Seattle, WA, USA.

* Correspondence: meren@uchicago.edu

Standfirst

Big data abounds in microbiology, but the workflows designed to enable researchers to interpret data can constrain the biological questions that can be asked. Five years after anvi'o was first published, this community-led multi-omics platform is maturing into an open software ecosystem that reduces constraints in 'omics data analyses.

Generating hundreds of millions of sequences from a microbial habitat is now commonplace for many microbiologists[1]. While the massive data streams offer detailed snapshots of the lifestyles of microorganisms, this data revolution in microbiology means that a new generation of computational tools is needed to empower life scientists in the era of multi-omics.

To meet the growing computational needs of the life sciences, computer scientists and bioinformaticians have created thousands of software tools[2]. These software fall into two general categories: 'essential tools' that implement functions fundamental to most bioinformatics tasks, and 'workflows' that make specific analytic strategies accessible.

If a comprehensive microbial 'omics investigation is a sophisticated dish, then essential tools are the kitchenware needed to cook. A chef can combine them in unique ways to answer any question, yet such freedom in data analysis not only requires the mastery of each essential tool but also demands experience in data wrangling and fluency in the command line environment to match the output format of one tool to the input requirements of the next. This barrier is overcome by workflows, which implement popular analysis strategies and make them accessible to those who have limited training in computation. If a comprehensive microbial 'omics investigation is a sophisticated dish, then each 'omics workflow is a recipe that turns raw material into a specific meal. For instance, a workflow for 'pangenomics' would typically take in a set of genomes and (1) identify open reading frames in all input genomes, (2) reciprocally align all translated amino acid sequences, (3) identify gene clusters by resolving pairwise sequence homology across all genes, and (4) report the distribution of gene clusters across genomes. By doing so, a software that implements pangenomics, such as Roary[3], would seamlessly run multiple essential tools consecutively, resolve input/output requirements of each, and address various ad hoc computational challenges to concoct a pangenome. Popular

efforts to make accessible workflows that form the backbone of 'omics-based microbiological studies include the Galaxy platform[4], bioBakery software collection[5], M-Tools (i.e., GroopM[6], CheckM[7]), and KBase[8]. While 'omics workflows conveniently summarise raw data into tables and figures, the ability to analyse data beyond pre-defined strategies they implement continues to be largely limited to master chefs, presenting the developers of 'omics workflows with a substantial responsibility: pre-determining the investigative routes their software enables users to traverse, which can influence how researchers interact with their data, conceivably affecting biological interpretations.

We introduced anvi'o (an analysis and visualisation platform for 'omics data) as an alternative solution for microbiologists who wanted more freedom in research questions they could ask of their data[9]. We started with what we regarded as the most pressing need at the time: a platform that enabled the reconstruction and interactive refinement of microbial genomes from environmental metagenomes. Fundamentals of this strategy were already established by those who pioneered genome-resolved metagenomics[10], but interactive visualisation and editing software that would enable microbiologists to intimately work with metagenome-assembled genomes was lacking. During the past five years anvi'o has become a community-driven software platform that currently stands upon more than 90,000 lines of open-source code and supports interactive and fully integrated access to state-of-the-art 'omics strategies including genomics, genome-resolved metagenomics and metatranscriptomics, pangenomics, metapangenomics, phylogenomics, and microbial population genetics (Figure 1).

Anvi'o differs from existing bioinformatics software due to its modular architecture, which enables flexibility, interactivity, reproducibility, and extensibility. To achieve this, the platform contains more than 100 interoperable programs, each of which performs individual tasks that can be combined to build new and unique analytical workflows. Anvi'o programs generate, modify, query, split, and merge anvi'o projects, which are really a set of extensible, self-contained SQLite databases. The interconnected nature of anvi'o programs which are glued together by these common data structures yields a network (http://merenlab.org/nt), rather than predetermined, linear paths for analysis. Through this modularity, anvi'o empowers its users to navigate through 'omics data without imposing rigid workflows.

Integrated interactive visualisation is at the center of anvi'o and helps researchers to engage with their data in all stages of analysis. Within the same interface, an anvi'o user can visualise amino acid sequence alignments between homologous genes across multiple genomes, investigate nucleotide-level coverage patterns and variants on the same DNA segment across metagenomes, interrogate associations between the genomic abundance and transcriptomic activity of environmental microbes, display phylogenetic trees and clustering dendrograms, and more. Furthermore, users can extend anvi'o displays with project-specific external data, increasing the utility of interactive interfaces for holistic descriptions of complex systems. The anvi'o interactive interface also provides its users with the artistic freedom to change colours, sizes, and drawing styles of display objects, add annotations, or reorder data layers for detailed communication of intricate observations. Because each anvi'o project is self-contained, researchers can easily make their analyses available online as a whole or in part, thereby enabling the integration, reusability, and reproducibility of their findings beyond static figures or tables. This strategy promotes transparency by permitting community validation and scrutiny through full access to data that underlie final conclusions.

Several key studies that used anvi'o during the past few years have demonstrated the integrative capabilities of the platform by implementing a combination of 'omics strategies to facilitate in-depth analysis of naturally occurring microbial habitats. For instance, Reveillaud and Bordenstein et al. reconstructed new genomes of *Wolbachia*, a fastidious endosymbiont[11], from individual insect ovary metagenomes, and computed a pangenome to compare these novel genomes to an existing reference[12]. They were then able to characterise the ecology of gene clusters in the environment by effectively combining metagenomics and pangenomics, discovering new members of the *Wolbachia* mobilome[12]. Yeoman et al. combined phylogenomics and pangenomics to infer ancestral relationships between a set of cultivar and metagenome-assembled genomes through a *de novo* identified set of single-copy core genes[13]. They demonstrated the correspondence among these genomes based on gene cluster membership patterns, phylogenomic inference, and average nucleotide identity in a single display[13]. Delmont and Kiefl et al. characterised the population structure of a subclade of SAR11, one of the most abundant microbial populations on Earth, by describing the

environmental core genes of a single genome across surface ocean metagenomes[14]. By linking single-amino acid variants in the environment to the predicted tertiary structures of these genes, they combined microbial population genetics with protein biochemistry to shed light on distinct evolutionary processes shaping the population structures of these bacteria[14]. Each of these studies employs unique approaches beyond well-established 'omics workflows to create rich, reproducible, and shareable data products (see http://merenlab.org/data).

Anvi'o does not implement strategies that take in raw data and produce summary tables or figures via a single command. As a result, anvi'o has a relatively steep learning curve. To address this, we have written extensive online tutorials that currently exceed 120,000 words, organised free workshops for hands-on anvi'o training, and created open educational resources to learn microbial 'omics. To interact with anvi'o users we set up an online forum and messaging service. During the past two years, more than 750 registered members of these services have engaged in technical and scientific discussions via more than 9,000 messages. But even when resources for learning are available, the journey from raw 'omics data to biological insights often takes a significant number of atomic steps of computation. To ameliorate the burden of scale and reproducibility in big data analyses we have also introduced anvi'o workflows, which automate routine computational steps of commonly used analytical strategies in microbial 'omics (http://merenlab.org/anvio-workflows). The anvi'o workflows are powered by Snakemake[15], which ensures relatively easy deployment to any computer system and automatic parallelisation of independent analysis steps. By turning raw input into data products to be analysed in the anvi'o software ecosystem, anvi'o workflows reduce the barriers for advanced use of computational resources and processing of large data streams for microbial 'omics.

As the developers of anvi'o who strive to create an open community resource, our next big challenge is to attract bioinformaticians to consider anvi'o as a software development ecosystem they can use for their own science. Any program that reads from or writes to anvi'o projects either directly (in any modern programming language) or through anvi'o application programmer interfaces (in Python) will immediately become accessible to anvi'o users, and

such applications will benefit from the data integration, interactive data visualisation, and error checking assurances anvi'o offers.

As an open-source platform that empowers microbiologists by offering them integrated yet uncharted means to steer through complex 'omics data, anvi'o welcomes its new users and contributors.

Acknowledgements

The URL https://github.com/merenlab/anvio/blob/master/AUTHORS.txt serves a complete list of anvi'o developers. We thank the creators of other open-source software tools for their generosity, anvi'o users for their patience with us, and Karen Lolans (0000-0003-1903-756X) for her critical reading of the manuscript and suggestions. The authors gratefully acknowledge support for anvi'o from the Simons Foundation and Alfred P. Sloan Foundation.

Author contributions

AME, EK, AS, IV, SEM, MSS, IF, JNP, MY, ECF, FT, ARW, OCE, RMM, QC, and ADW coded and documented anvi'o, contributed to the implementation of new analytical strategies, and engaged with the anvi'o community. MDL, VK, EDG, BDM, and AK wrote blog posts and tutorials to make anvi'o accessible to the broader community. XVC, LJM helped with technical issues and testing of new features on GitHub. EAM, SLRS, and RA created undergraduate and graduate-level educational material and taught anvi'o. LM organized workshops for the training of research professionals. JF, AFG, LM, TOD, and ADW made intellectual contributions that influenced the direction of the platform. AME wrote the paper and prepared the figure with input from all authors.

Competing interests

Authors have no conflicts of interest to declare.

Figure Legends

Figure 1. A glimpse of the interconnected nature of ’omics analysis strategies anvi’o makes accessible, and their potential applications.

References

1. White, R. A., Callister, S. J., Moore, R. J., Baker, E. S. & Jansson, J. K. The past, present and future of microbiome analyses. *Nat. Protoc.* **11**, 2049–2053 (2016).
2. Callahan, A., Winnenburg, R. & Shah, N. H. U-Index, a dataset and an impact metric for informatics tools and databases. *Sci Data* **5**, 180043 (2018).
3. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
4. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).
5. McIver, L. J. *et al.* bioBakery: a meta’omic analysis environment. *Bioinformatics* vol. 34 1235–1237 (2018).
6. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
7. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
8. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).
9. Eren, A. M. *et al.* Anvi’o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**,

e1319 (2015).

10. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

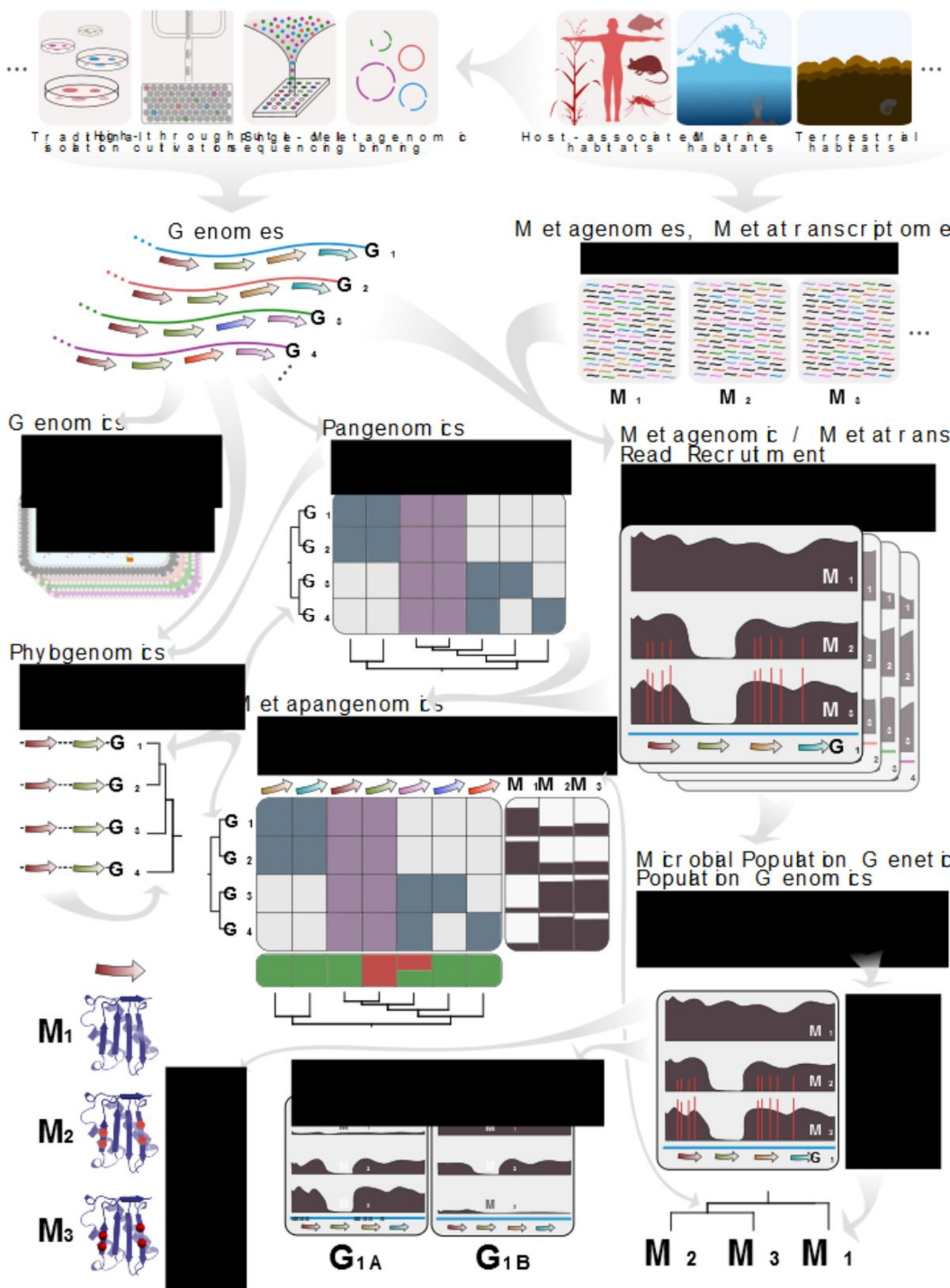
11. Werren, J. H., Baldo, L. & Clark, M. E. Wolbachia: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* **6**, 741–751 (2008).

12. Reveillaud, J. *et al.* The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat. Commun.* **10**, 1051 (2019).

13. Yeoman, C. J. *et al.* Genome-resolved insights into a novel Spiroplasma symbiont of the Wheat Stem Sawfly (Cephus cinctus). *PeerJ* **7**, e7548 (2019).

14. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* **8**, (2019).

15. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).


Traditional isolation
High-throughput cultivation
Single-cell sequencing
Metagenomic binning
Host-associated habitats
Marine habitats
Terrestrial habitats
Genomes
Metagenomes, Metatranscriptomes
Genomics
Pangenomics
Metagenomic / Metatrans
Read Recruitment
Phybgenomics
Metapangenomics
Microbial Population Genetics
Population Genomics
G1A
G1B