

<https://helda.helsinki.fi>

---

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

Eren, A. Murat

2021-01

---

Eren , A M , Kiefl , E , Shaiber , A , Veseli , I , Miller , S E , Schechter , M S , Fink , I , Pan , J N , Yousef , M , Fogarty , E C , Trigodet , F , Watson , A R , Esen , O C , Moore , R M , Clayssen , Q , Lee , M D , Kivenson , V , Graham , E D , Merrill , B D , Karkman , A , Blankenberg , D , Eppley , J M , Sjodin , A , Scott , J J , Vazquez-Campos , X , McKay , L J , McDaniel , E A , Stevens , S L R , Anderson , R E , Fuessel , J , Fernandez-Guerra , A , Maignien , L , Delmont , T O & Willis , A D 2021 , ' Community-led, integrated, reproducible multi-omics with anvi'o ' , Nature Microbiology , vol. 6 , pp. 3-6 . <https://doi.org/10.1038/s41564-020-00834-3>

---

<http://hdl.handle.net/10138/332078>

<https://doi.org/10.1038/s41564-020-00834-3>

---

acceptedVersion

---

*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

# 1 Community-led, integrated, reproducible multi-omics with anvio

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

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

36 \* Correspondence: [meren@uchicago.edu](mailto:meren@uchicago.edu)

37

38 **Standfirst**

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

43

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

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

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

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

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

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

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

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

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

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

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

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

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

## 158 Acknowledgements

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

## 164 Author contributions

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

## 174 Competing interests

175 Authors have no conflicts of interest to declare.

## 176 Figure Legends

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

## 179 References

- 180 1. White, R. A., Callister, S. J., Moore, R. J., Baker, E. S. & Jansson, J. K. The past, present and future of  
181 microbiome analyses. *Nat. Protoc.* **11**, 2049–2053 (2016).
- 182 2. Callahan, A., Winnenburg, R. & Shah, N. H. U-Index, a dataset and an impact metric for informatics  
183 tools and databases. *Sci Data* **5**, 180043 (2018).
- 184 3. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–  
185 3693 (2015).
- 186 4. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical  
187 analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).
- 188 5. McIver, L. J. *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* vol. 34 1235–1237  
189 (2018).
- 190 6. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from  
191 related metagenomes. *PeerJ* **2**, e603 (2014).
- 192 7. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the  
193 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*  
194 **25**, 1043–1055 (2015).
- 195 8. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase.  
196 *Nat. Biotechnol.* **36**, 566–569 (2018).
- 197 9. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**,



- 198 e1319 (2015).
- 199 10. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial  
200 genomes from the environment. *Nature* **428**, 37–43 (2004).
- 201 11. Werren, J. H., Baldo, L. & Clark, M. E. Wolbachia: master manipulators of invertebrate biology. *Nat.*  
202 *Rev. Microbiol.* **6**, 741–751 (2008).
- 203 12. Reveillaud, J. *et al.* The Wolbachia mobilome in *Culex pipiens* includes a putative plasmid. *Nat.*  
204 *Commun.* **10**, 1051 (2019).
- 205 13. Yeoman, C. J. *et al.* Genome-resolved insights into a novel Spiroplasma symbiont of the Wheat  
206 Stem Sawfly (*Cephus cinctus*). *PeerJ* **7**, e7548 (2019).
- 207 14. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the  
208 biogeography of a global SAR11 subclade. *Elife* **8**, (2019).
- 209 15. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**,  
210 2520–2522 (2012).

