AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE

Using Cartesian Doubt To Build a Sequencing-Based View of Microbiology

Braden T. Tierney,[a,b,c,d] Erika Szymanski,[e] James R. Henriksen,[f] Aleksandar D. Kostic,[b,c,d] Chirag J. Patel[a]

[a]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA
[b]Section on Pathophysiology and Molecular Pharmacology, Joslin Diabetes Center, Boston, Massachusetts, USA
[c]Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center, Boston, Massachusetts, USA
[d]Department of Microbiology, Harvard Medical School, Boston, Massachusetts, USA
[e]Department of English, Colorado State University, Fort Collins, Colorado, USA
[f]Mr. Fusion Inc., Fort Collins, Colorado, USA

ABSTRACT The technological leap of DNA sequencing generated a tension between modern metagenomics and historical microbiology. We are forcibly harmonizing the output of a modern tool with centuries of experimental knowledge derived from culture-based microbiology. As a thought experiment, we borrow the notion of Cartesian doubt from philosopher Rene Descartes, who used doubt to build a philosophical framework from his incorrigible statement that "I think therefore I am." We aim to cast away preconceived notions and conceptualize microorganisms through the lens of metagenomic sequencing alone. Specifically, we propose funding and building analysis and engineering methods that neither search for nor rely on the assumption of independent genomes bound by lipid barriers containing discrete functional roles and taxonomies. We propose that a view of microbial communities based in sequencing will engender novel insights into metagenomic structure and may capture functional biology not reflected within the current paradigm.

KEYWORDS Cartesian doubt, microbial genetics, microbial species concept, microbiome

Cartesian doubt—beginning with radical skepticism and moving forward with as few external assumptions as possible—can be used to reconceive our approaches to microbiome science, potentially avoiding biases and conflicts stemming from centuries of culture-based microbiology. In 1641, Rene Descartes published his *Meditations on First Philosophy*, in which he upended and tossed aside past philosophical thought by asking "How do we know what is true?" (1). We propose similarly rethinking microbial communities as revealed via DNA sequencing, reimagining what microbial life may be instead of assuming what it is based on existing understandings of taxonomy, microbial genomes, or other culture-centric paradigms.

Consider metagenomic sequencing as an incarnation of Anton van Leeuwenhoek's microscope: peering through its "lens," what do we "see?" A FASTA file certainly does not display the discrete particles Anton van Leeuwenhoek described: sequencing is a lens foreign to historical microbiologists' view of microbes. Nevertheless, microbiome science routinely maps sequencing reads to "species" and "core" or "accessory" genes. Why restrict metagenomes to this paradigm, overlaying modern tools with centuries of single-species-centric experimental work rooted in physical observation? (2, 3). In light of the potential for epistemic conflicts between culture-based and sequencing-based knowledge, can the field establish an analytic frame that integrates these distinct perspectives?

Gaps between metagenomics (4) and historical microbiology illustrate why microbiome scientists should reconsider our core assumptions (Fig. 1A)—though the field

Citation Tierney BT, Szymanski E, Henriksen JR, Kostic AD, Patel CJ. 2021. Using Cartesian doubt to build a sequencing-based view of microbiology. mSystems 6:e00574-21. https://doi.org/10.1128/mSystems.00574-21.

Editor Linda Kinkel, University of Minnesota

Address correspondence to Braden T. Tierney, btierney@g.harvard.edu, Aleksandar D. Kostic, Aleksandar.Kostic@joslin.harvard.edu, or Chirag J. Patel, chirag_patel@hms.harvard.edu.

Received 7 May 2021
Accepted 23 September 2021
Published 12 October 2021

should not adopt an ahistorical view. Rather, researchers should acknowledge that contemporary analyses can be biased by prior experiments. For example, phylogenies (5) employed as buckets for sequence data amalgamate physiological (e.g., via Bergey's manual and numerical taxonomy) and genetic markers (e.g., 16S sequence similarity) built through specific, now-historical perceptions of microbial life (6–10). This can constrain our understanding of microbiome biology (Fig. 1B). Additionally, "complete" genomes are defined through gold standard cutoffs that prioritize genes on the basis of their presence in previously assembled sequence data (11). Ecosystem-spanning signals extraneous to our current frameworks, like horizontal gene transfer (HGT) or evolutionary drift, look like noise to a framework built for monocultures and not communities (Fig. 1C). Assembly-based methods for genome discovery may therefore artificially bias gene content in organisms—or functions—with high rates of HGT. Further, the functional roles of similar sequences are often defined through global percent identity cutoffs, despite sequence not necessarily correlating to function (12). Finally, bio- and geochemical reactions exist in multiple spatial and temporal structures that may not be membrane bound within discrete cells. Overall, microbiologists constructed paradigms to cohere with pure cultures; a sequencing-centered approach to metagenomics unconstrained by pure-culture-based paradigms provides an exciting opportunity to rethink assumptions about the organization of microbial life.

While it is impossible to truly disregard a preconceived framework derived from hundreds of years of experimentation, Cartesian doubt can address epistemological conflicts between observations (i.e., raw data) from microbial communities and the paradigms (i.e., theory) used to interpret them. Consider working from the following axiom: a metagenome is captured in a data structure representing complete "sequencing of microbial DNA"—base pair order (e.g., reads), chemical structure (e.g., methylation), and spatial structure (e.g., via Hi-C [high-throughput chromosome conformation capture]). In other words, we hypothesize that DNA sequencing will advance such that it operates at any read length with increased resolution for sequence chemistry and structure. An unprejudiced view of this idealized sequencing data would allow the field to, at least temporarily, abnegate the paradigms that currently bind us and identify novel metagenomic structure.

Microbiome pattern identification is initially an algorithmic task. Modern approaches to metagenomic data analysis today discard ostensibly junk reads that, for example, do not map to draft genomes or assemble cleanly; unbiased approaches should first aim to minimize discarded data to avoid biological signal loss (13–16). Methodologically, numerous tools are used to "project" complex, unordered data into human-readable, low-dimensional space (Fig. 2A). These tools stem mostly from computer science and natural language processing, and some have already been applied to metagenomic data (17, 18). One simple method may be to collect k-mers in individual sequence reads across time, collapsing them into highly correlated clusters. Researchers could also consider using extensions of vector-based sequence projection methods, colored de Bruijn graphs, or metabolic network strategies (19–24). These approaches will advance analyses unconstrained by the paradigm of individual cells containing individual genomes. However, methods and data structures (e.g., databases indexing the k-mers of the Sequence Read Archive) should be selected carefully, as different questions mandate different approaches.

Would a sequence-first approach revise our view of genomes, genes, or codons? Any algorithm effectively parsing read data will identify conservation in sequence. Consider an approach that identifies consistent patterns in DNA base pair order. This may identify codons, as they are conserved and nonrandom. Comparison across reads could uncover alternate coding schemes as variations within this pattern (25). Perhaps longer reads would recover genes with little sequence divergence. Genes with high divergence would likely not cocluster; however, conserved motifs may. Biologists might have to further reconsider the fundamental units of microbial genetics (26, 27)

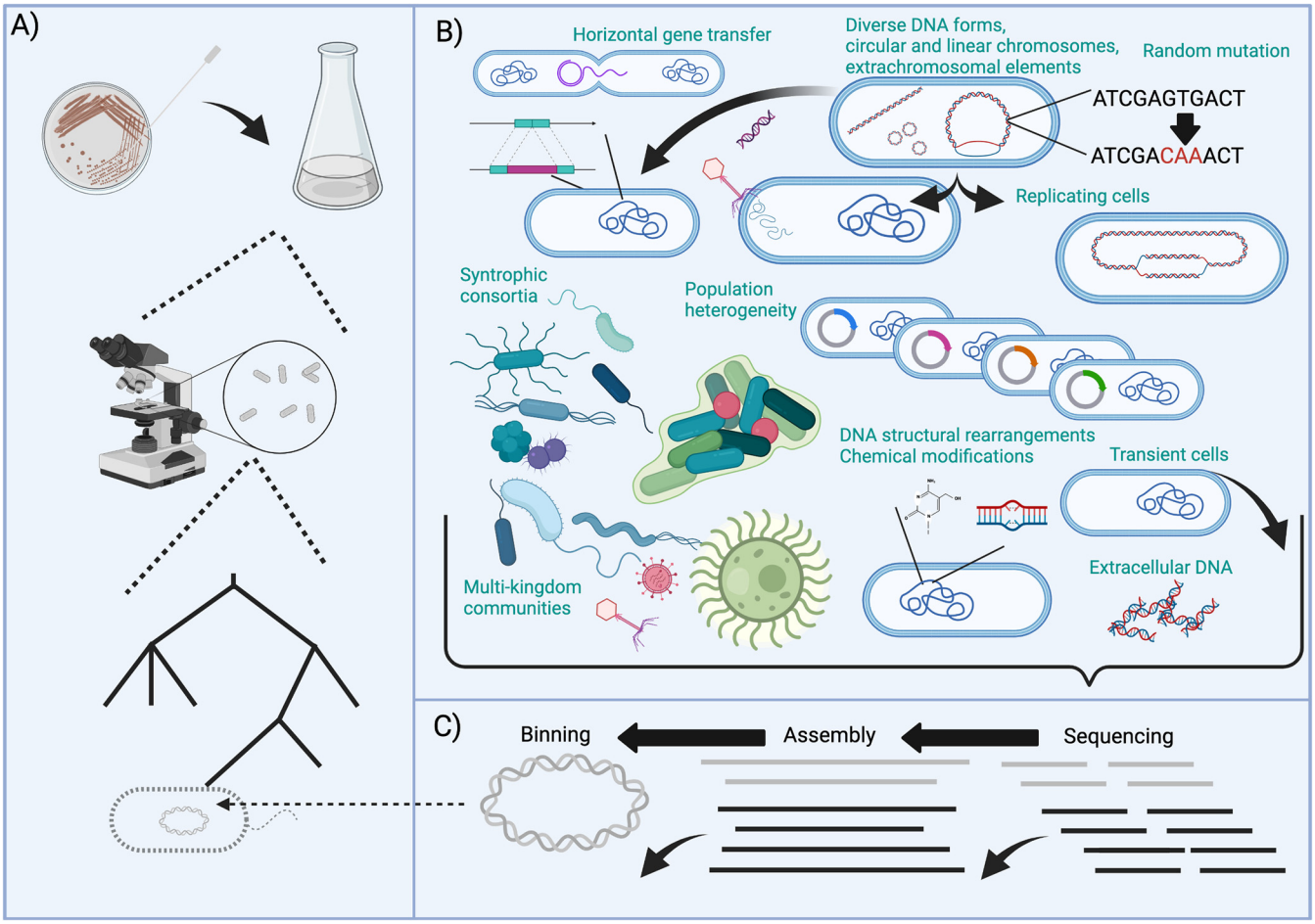

FIG 1 The existing paradigm of microbiome science. (A) Our historical view of microbes originates from what is culturable. Bacteria, specifically, have been mostly observed in clonal isolation and are assumed to have measurable cell-based genomes that can be hierarchically grouped by phylogenetics. (B) Microbiome scientists (generally) use DNA sequencing to investigate a complex, multikingdom microbial community that is changing across space and time through a series of complex interactions that are not well represented by this framework, including horizontal gene transfer, cell replication, and spontaneous mutation. (C) To build a sense of microbial (bacterial in this case) genomes, researchers, for example, assemble sequencing reads into contigs and bin contigs into "complete," phylogenetically annotated, genomes. The figure was generated with BioRender.com.

For example, would core and accessory genes—or other patterns entirely—exist at higher levels of genomic organization (e.g., across metagenomes instead of genomes)?

Analyses based upon existing paradigms may also be limited in their capacities to capture genomic temporal variation. No microbial genome (or genome within an organism in any kingdom of life) is static across time and space. Replication forks, structural rearrangements, CRISPR spacer acquisition and loss, HGT, and plasmids will yield continually "incomplete" genomes, even if a single read could capture a contiguous unit of DNA. Unbiased pattern identification that aims to discover fundamental, spatiotemporally consistent (or inconsistent) metagenomic units will align our view of microbial genomics along an entirely new axis, redefining our perspective on microbiomic temporal modulation (Fig. 2B).

We hypothesize that unbiased approaches to sequence analysis would yield a continuum of sequence-based conservation: sequence substructures (e.g., motifs, genes conserved at high percent identity) that represent emergent biology, not necessarily tied to pathways, genes, or genomes. These substructures could, however, be periodically cooccurring, dynamic (or temporally periodic) elements that may, for example, be environmentally dependent or affect ecosystem-level functions. This "periodic table of metagenomic elements," which would minimize assumptions about meaningful versus noise reads, could provide increased insight into latent metagenomic structure.

Historically, important biology has been overlooked (e.g., the kingdom of archaea,

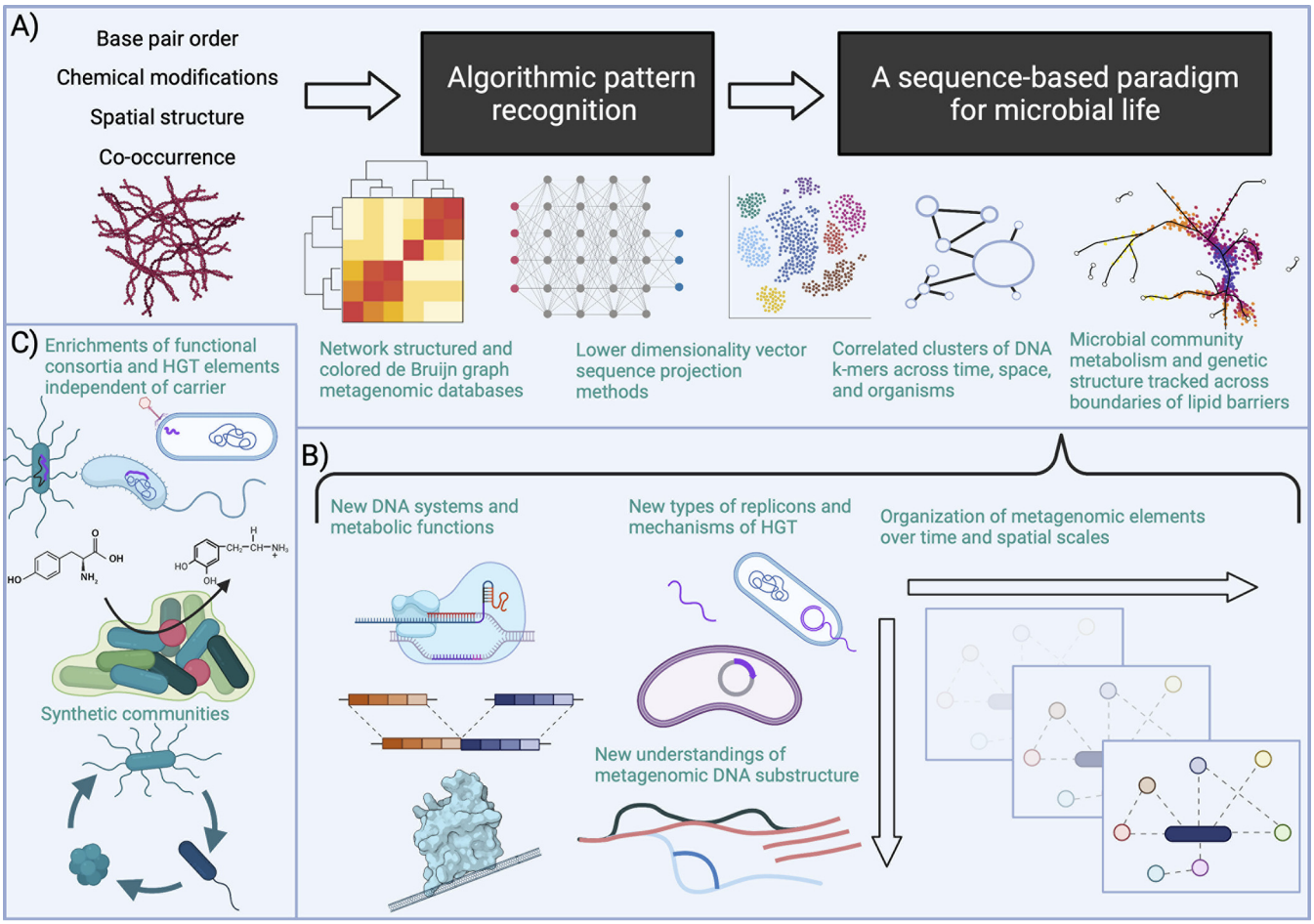

FIG 2 Discovering new frameworks with Cartesian doubt. We propose using Cartesian doubt to consider sequencing data (referring to a range of multi-omic technologies) and how unbiased pattern recognition (A) can result in a cell-agnostic, sequencing-based paradigm that would be complementary to but unguided by the experimental history of microbiology (B). Combined with novel wet-lab techniques working within this new view of metagenomics, microbiome scientists could thereby reveal potentially unknown biology outside the scope of our current framework (C). The figure was generated with BioRender.com.

noncoding DNA, noncanonical amino acids) because technology was not designed to detect it or because assumptions limited the capacity to interpret biological signals, even at times construing such signal as contamination (28, 29). The reads that float between disparate genomes (or nodes that cannot be resolved in *de novo* assemblies) should be treated as signals, not hidden by forcing resolution or filtered out by data handling. Hundreds of thousands of reference microbial genomes derive from metagenome-assembled genomes (MAGs) built using culture-based "gold standards," which may exclude genes (e.g., conjugation systems) (30) or may amplify genomic features like random mutation, HGT, or doubling under the guise of new genome discovery. Finally, considering beyond base pair order-focused approaches will facilitate the incorporation of alternative sequencing data (e.g., Hi-C) into our understanding of metagenomic communities, their meta-phenotypes (e.g., colonization resistance), and their diversity (e.g., bacteria, fungi, and viruses).

Applying Cartesian doubt to microbiome science has numerous applications to rethinking the rules of life for microbiomes, ranging from our view of metagenomic DNA substructure to the microbial species concept to the tools used to work with metagenomes. We challenge the scientific and funding communities to pursue three efforts in particular. First, since microbial metabolism is not bounded by lipid barriers in a community setting, neither should our metagenomic paradigm. Scientists need to extend (20, 22, 23, 31, 32) and create new algorithms that integrate across sequencing

modalities and consider metagenomes as greater than sums of their parts. Second, theory and empirical data collection (i.e., experimental practice) need to inform each other. Currently, the field's assumptions constrain methodological development. If microbiological theory were less historically biased, further (33–35) wet-lab techniques for operating on different units of microbial life could be developed, perhaps extending on current synthetic community work but relying more on enrichments of functional consortia or independent HGT elements (Fig. 2C). Data interpretation methods are also needed, such as theoretical modeling (20) and algorithms operating on k-mers and microbiome metabolism. Finally, "gold standards" must be defined only in the context of a particular research question, avoiding claims regarding universality, as doing so obscures assumptions that may be invalid in context (e.g., >95% sequence identity when comparing genes or "complete" genomes).

Minimizing assumptions will add complementary insight to current paradigms while adding richness to our understanding of the functional organization of complex microbiomes. Indeed, Cartesian doubt's true power is accommodating many different perspectives, not necessarily unveiling some grand truth, but rather adjusting reference points through complementary scientific lenses. The historical model of microbiology has gotten us extremely far, and its value cannot be overlooked. However, while we all stand on the shoulders of giants, it is occasionally prudent to consider the ground beneath our feet.

REFERENCES

1. Descartes R. 2008. Meditations on first philosophy: with selections from the objections and replies. Oxford University Press, Oxford, United Kingdom.
2. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
3. Segerman B. 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol 2:116. https://doi.org/10.3389/fcimb.2012.00116.
4. Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004.
5. Sanford RA, Lloyd KG, Konstantinidis KT, Löffler FE. 2021. Microbial taxonomy run amok. Trends Microbiol 29:394–404. https://doi.org/10.1016/j.tim.2020.12.010.
6. Sneath PHA. 2005. Numerical taxonomy, p 39–42. *In* Brenner DJ, Krieg NR, Staley JT, Garrity GM (ed), Bergey's manual of systematic bacteriology, vol 2. The Proteobacteria. Part A introductory essays. Springer US, Boston, MA.
7. Murray RGE, Holt JG. 2005. The history of Bergey's manual, p 1–14. *In* Brenner DJ, Krieg NR, Staley JT, Garrity GM (ed), Bergey's manual of systematic bacteriology, vol 2. The Proteobacteria. Part A introductory essays. Springer US, Boston, MA.
8. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090. https://doi.org/10.1073/pnas.74.11.5088.
9. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, Truper HG, Murray RGE, Grimont PAD, Brenner DJ, Starr MP, Moore LH. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol 37:463–464. https://doi.org/10.1099/00207713-37-4-463.
10. Kang C-H, Nam Y-D, Chung W-H, Quan Z-X, Park Y-H, Park S-J, Desmone R, Wan X-F, Rhee S-K. 2007. Relationship between genome similarity and DNA-DNA hybridization among closely related bacteria. J Microbiol Biotechnol 17:945–951.
11. Hugenholtz P, Skarshewski A, Parks DH. 2016. Genome-based microbial taxonomy coming of age. Cold Spring Harb Perspect Biol 8:a018085. https://doi.org/10.1101/cshperspect.a018085.
12. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65. https://doi.org/10.1038/nature08821.
13. Maximilian O, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C-C, Mueller KA, Sullivan ST, Chain PSG, Liachko I. 2017. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. bioRxiv. https://doi.org/10.1101/198713.
14. Grassl N, Kulak NA, Pichler G, Geyer PE, Jung J, Schubert S, Sinitcyn P, Cox J, Mann M. 2016. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome Med 8:44. https://doi.org/10.1186/s13073-016-0293-0.
15. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature 550:345–353. https://doi.org/10.1038/nature24286.
16. Ma Q, Bücking H, Gonzalez Hernandez JL, Subramanian S. 2019. Single-cell RNA sequencing of plant-associated bacterial communities. Front Microbiol 10:2452. https://doi.org/10.3389/fmicb.2019.02452.
17. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, Dekas AE, Batzoglou S, Bhatt AS. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat Biotechnol 36:1067–1075. https://doi.org/10.1038/nbt.4266.
18. Menegaux R, Vert J-P. 2019. Continuous embeddings of DNA sequencing reads and application to metagenomics. J Comput Biol 26:509–518. https://doi.org/10.1089/cmb.2018.0174.
19. Weber M, Teeling H, Huang S, Waldmann J, Kassabgy M, Fuchs BM, Klindworth A, Klockow C, Wichels A, Gerdts G, Amann R, Glöckner FO. 2011. Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. ISME J 5:918–928. https://doi.org/10.1038/ismej.2010.180.
20. Cai J, Tan T, Chan SHJ. 2021. Predicting Nash equilibria for microbial metabolic interactions. Bioinformatics 36:5649–5655. https://doi.org/10.1093/bioinformatics/btaa1014.
21. Biggs MB, Papin JA. 2016. Metabolic network-guided binning of metagenomic sequence fragments. Bioinformatics 32:867–874. https://doi.org/10.1093/bioinformatics/btv671.
22. Yau SS-T, Wang J, Niknejad A, Lu C, Jin N, Ho Y-K. 2003. DNA sequence representation without degeneracy. Nucleic Acids Res 31:3078–3080. https://doi.org/10.1093/nar/gkg432.
23. Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. Genome Res 27:665–676. https://doi.org/10.1101/gr.214155.116.

24. Titus BC, Moritz D, O'Brien MP, Reidl F, Reiter T, Sullivan BD. 2019. Exploring neighborhoods in large metagenome assembly graphs reveals hidden sequence diversity. bioRxiv. https://doi.org/10.1101/462788.
25. Shulgina Y, Eddy SR. 2021. A computational screen for alternative genetic codes in over 250,000 genomes. bioRxiv. https://doi.org/10.1101/2021.06.18.448887.
26. Keller EF. 2011. Genes, genomes, and genomics. Biol Theory 6:132–140. https://doi.org/10.1007/s13752-012-0014-x.
27. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681. https://doi.org/10.1101/gr.6339607.
28. Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, Ingram J, Nene RV, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756. https://doi.org/10.1126/science.1142490.
29. Al-Shayeb B, Schoelmerich MC, West-Roberts J, Valentin-Alvarado LE, Sachdeva R, Mullen S, Crits-Christoph A, Wilkins MJ, Williams KH, Doudna JA, Banfield JF. 2021. Borgs are giant extrachromosomal elements with the potential to augment methane oxidation. bioRxiv. https://doi.org/10.1101/2021.07.10.451761.
30. Joris BR, Browne TS, Hamilton TA, Edgell DR, Gloor GB. 2021. Separation of cohorts on the basis of bacterial type IV conjugation systems identified from metagenomic assemblies. bioRxiv. https://doi.org/10.1101/2021.04.15.440092.
31. Forsdyke DR. 2019. Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny. arXiv 1903.04866 [q-bioPE].
32. Abdelkareem AO, Khalil MI, Elbehery AHA, Abbas HM. 2020. Viral sequence identification in metagenomes using natural language processing techniques. bioRxiv. https://doi.org/10.1101/2020.01.10.892158.
33. Cira NJ, Pearce MT, Quake SR. 2018. Neutral and selective dynamics in a synthetic microbial community. Proc Natl Acad Sci U S A 115:E9842–E9848. https://doi.org/10.1073/pnas.1808118115.
34. Pacheco AR, Osborne ML, Segrè D. 2021. Non-additive microbial community responses to environmental complexity. Nat Commun 12:2365. https://doi.org/10.1038/s41467-021-22426-3.
35. Zhu W, Winter MG, Byndloss MX, Spiga L, Duerkop BA, Hughes ER, Büttner L, de Lima Romão E, Behrendt CL, Lopez CA, Sifuentes-Dominguez L, Huff-Hardy K, Wilson RP, Gillis CC, Tükel Ç, Koh AY, Burstein E, Hooper LV, Bäumler AJ, Winter SE. 2018. Precision editing of the gut microbiota ameliorates colitis. Nature 553:208–211. https://doi.org/10.1038/nature25172.