





Embracing Metagenomic Complexity with a Genome-Free Approach

 Izaak Coleman,^a  Tal Korem^{a,b,c}

^aProgram for Mathematical Genomics, Department of Systems Biology, Columbia University Irving Medical Center, New York, New York, USA

^bDepartment of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, New York, USA

^cCIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada

ABSTRACT A central paradigm in microbiome data analysis, which we term the genome-centric paradigm, is that a linear (non-branching) DNA sequence is the ideal representation of a microbial genome. This representation is natural, as microbes indeed have non-branching genomes. Tremendous discoveries in microbiology were made under this paradigm, but is it always optimal for microbiome research? In this Commentary, we claim that the realization of this paradigm in metagenomic assembly, a fundamental step in the “metagenomics analysis pipeline,” suboptimally models the extensive genomic variability present in the microbiome. We outline our efforts to address these issues with a “genome-free” approach that eschews linear genomic representations in favor of a pan-metagenomic graph.

KEYWORDS assembly, genomics, metagenomics, microbiome


Microbiomes often contain hundreds of species, with a highly complex metagenomic structure; even distantly related microbes share genomic material (1, 2) due to vertical inheritance and horizontal transfers, and even closely related strains diverge (2–4). Variation is present within and between microbiomes (2–5), and occurs over relatively short timescales (4, 6). Understanding this variability is critical for topics such as emergence and maintenance of antibiotic resistance (7), in which horizontal gene transfer plays an important role (8). It has also been associated with host phenotypes (2–6), pinpointing specific genomic regions that are potentially adaptive to a particular host. In a recent study, we showed that a functional analysis of variable regions can even offer mechanistic hypotheses explaining such associations (2).

Variable genomic regions are likely poorly represented in reference genomes. Reference genomes are assembled from different populations, clinical conditions, or habitats, and have therefore been exposed to different environments and selective pressures. This means that they, and the variable genomic regions they encode, are likely irrelevant to the samples under study. A major promise therefore lies in *de novo* assembly, which directly models all the information present in a metagenomic sample. Recent studies, however, have demonstrated that state-of-the-art assemblers work well mostly for highly abundant strains with low heterogeneity (9), and are depleted of critical components such as mobile genetic elements (10). Here, we claim this is a direct result of the genome-centric paradigm. We argue for a “genome-free” approach, which does not attempt to produce linear assemblies but instead uses a “pan-metagenomic” graph (Fig. 1) that directly represents genomic variability across microbes in multiple samples. While we focus on analysis of short-read sequencing data, similar arguments could be made for long-read data. Our belief is that this approach offers a better framework for studying genomic variability in the microbiome.

Citation Coleman I, Korem T. 2021. Embracing metagenomic complexity with a genome-free approach. *mSystems* 6:e00816-21. <https://doi.org/10.1128/mSystems.00816-21>.

Copyright © 2021 Coleman and Korem. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Tal Korem, tal.korem@columbia.edu.

 Is a linear reference genome the ideal representation of a microbial genome? In their commentary, Coleman & Korem claim that a “genome-free”, graph-based representation of the metagenome could be key to understanding genomic variability in the microbiome.

Conflict of Interest Disclosures: I.C. has nothing to disclose. T.K. has nothing to disclose.

The views expressed in this article do not necessarily reflect the views of the journal or of ASM.

This article is part of a special series sponsored by Floré.

Published 17 August 2021

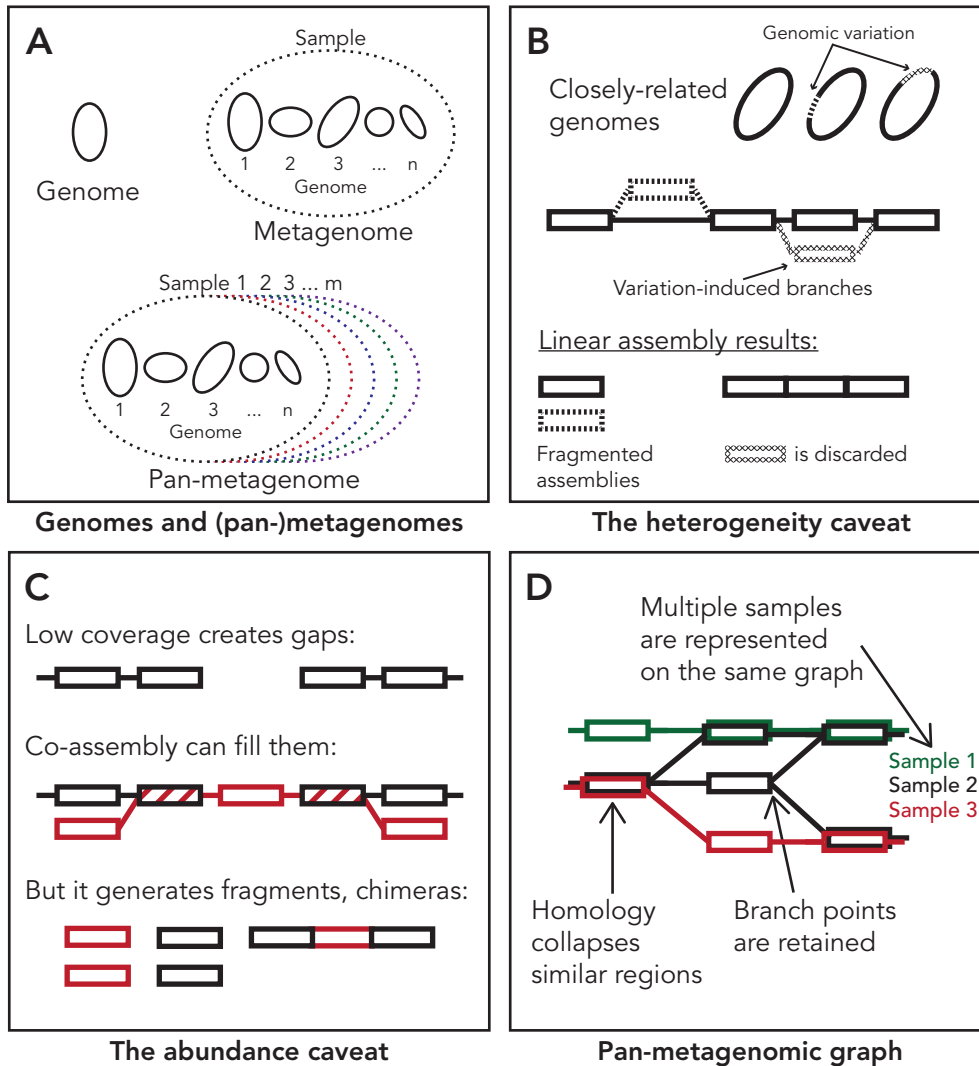


FIG 1 (A) Visual comparison between a genome; a metagenome, the collection of all genomes from a sampled microbial community; and a pan-metagenome, a collection of genomes, each deriving from one of multiple sampled communities. (B) The heterogeneity caveat: genomic variation between closely-related genomes (dashed sections) induces branching structures in assembly graphs (dashed nodes and edges). Linear assembly breaks down these structures, resulting in either fragmented contigs or the removal of variable regions. (C) The abundance caveat: undersampling of low-abundance genomes creates gaps in their assemblies. Co-assembly attempts to exploit information from close-matching genomes in other samples (red path) to fill these gaps. Some regions from these genomes are identical (diagonally striped nodes) and facilitate co-assembly; others are divergent, and introduce additional branching to the graph. This may result in either chimeras or fragmented contigs, and lower-quality assemblies in general. (D) We propose a graph-based representation of the pan-metagenome that addresses the caveats of the current paradigm. Our representation models metagenomic data across multiple samples, while keeping track of the originating sample of each sequence (red, black, and green). Sequence homology is used to collapse similar genomic regions (overlapping nodes), attenuating excessive branching within the graph in order to reveal variation at different scales with no information loss.

THE GENOME-CENTRIC PARADIGM FAILS TO CAPTURE THE PAN-METAGENOME

Contemporary assemblers (11–14) follow a similar process that realizes the genome-centric paradigm: sequencing reads are tiled into an assembly graph, which is then traversed to find paths representing linear contigs supported by the data. The goal of these assemblers is to generate the longest linear contigs possible, as reflected in some of the metrics used to assess assembly quality, such as N50.

Generating linear contigs is done at the cost of disregarding variation. When an assembler reaches a variation-induced branching structure in the graph (Fig. 1B), either one branch is selected over the other using some heuristic, such as removal of low-

abundance variants that are assumed to originate in sequencing errors, or the branching structure is broken into multiple non-branching contigs (15, 16). In either case, the information contained in branched structures, which directly represents variability, is lost for downstream analyses. Indeed, assemblies of heterogeneous strains are typically poor in quality (9), likely due to sequence heterogeneity creating complex branching topology that assemblers cannot resolve, and instead fragment. This heterogeneity caveat of disregarding variation has a major impact on mobile elements and horizontally transferred genes, which are typically depleted from assemblies (Fig. 1B) (10).

Albeit less directly, the genome-centric paradigm also affects the assembly of low-abundance strains. A recent large-scale study demonstrated that high-quality metagenome-assembled genomes are generated only for genomes with approximately 10 to 20 \times coverage, attainable only for the most abundant strains in each sample (9). It is likely that strains with lower abundance simply lack the coverage that will facilitate a high-quality assembly from a single sample. This issue could be addressed by using information from closely-related strains present in other samples, an approach termed “co-assembly.” Co-assembly, however, also introduces additional complexity to the assembly graph, generating branches representing heterogeneity and homology between similar strains from different samples. As with the heterogeneity caveat, assemblers typically break these branches, resulting in fragmented contigs. In some cases, they might even traverse paths through them, introducing chimeras—contigs composed of multiple different strains. Consequently, co-assembly under the genome-centric paradigm reduces the quality of assemblies (17) and is not commonly used in our field (9, 18, 19). We refer to this effect of the genome-centric paradigm on assemblers as the abundance caveat (Fig. 1C).

In summary, the realization of the genome-centric paradigm in metagenomic assembly results in a suboptimal representation of the variability across microbiomes, particularly evident in low-abundance and heterogenous strains. At the heart of both the abundance and heterogeneity caveats is the fact that to comply with the genome-centric paradigm, and generate linear contigs, assemblers need to resolve branching structures. These structures, however, directly encode the genomic variability that we are interested in. It is not a surprise, then, that some reference-based methods attempt to detect exactly these branching structures by analyzing clipped read-mappings or variations in read-coverage (2, 20). We propose a more direct approach.

BEYOND GENOMES: MODELING THE PAN-METAGENOME

In order to model variability within and across microbiomes, we are shifting our analytic representation of metagenomic data away from the genome-centric paradigm, toward the non-linear graph-based representation of the pan-metagenome: the entire collection of genetic elements present across multiple metagenomes (Fig. 1A). We use this representation to better model genomic variability in the microbiome, retaining the non-linear branching structures that encode variability. Being pan-metagenomic, our graph jointly models data from multiple samples. The originating samples of each sequence are recorded, facilitating comparative analyses. As we detail below, our framework also addresses the heterogeneity and abundance caveats (Fig. 1D), and could form the basis for extensive downstream analyses.

Heterogeneity induces a complex and nested branching structure in the pan-metagenome. Single nucleotide polymorphisms (SNPs) and small indels occur within larger structural variants, which may themselves show internal repetitive structure or homology to other genomes. In an attempt to construct long, linear contigs, assemblers resolve these branching structures; the consequential loss of variant information is the heterogeneity caveat. While we want to retain these branching structures and the information they encode, we also wish to control and attenuate the complexity of the resulting topology, in order to facilitate downstream analyses. We therefore use sequence homology to determine when branching should occur: sequences that are homologous according to a user-defined threshold are joined together, providing

control over the topological complexity of the graph, without falling to the heterogeneity caveat. A similar approach was recently applied to long-read assembly (21). Our ability to simplify topology allows us to reveal the large-scale structural architecture of the pan-metagenome without losing fine-scale variation.

By combining information on closely-related strains across samples, co-assembly could improve the genomic information modeled for each strain. At the same time, it introduces additional complexity in the form of branched structures. As described for the abundance caveat, current assemblers either break down these structures or traverse chimeric paths through them. We approach this problem differently. Whereas current co-assembly approaches operate “blindly,” without utilizing information about the originating sample of each read, we use “informed co-assembly,” which exploits both this information and information about the genome sequence recoverable from each sample. This allows us to intentionally introduce chimeras when we believe that, based on recoverable sequence, two strains from different samples are similar enough such that a gap in one can be filled with sequence from the other. At the same time, we are able to ignore branching structures representing homology between distant strains, as if assembly within these regions was performed in a sample-specific manner. Consequently, informed co-assembly within our framework mitigates the adverse properties of co-assembly under the genome-centric paradigm. We flag the chimeras we introduce, enabling flexible and informed use of chimeras by downstream analyses.

Embracing non-linearity facilitates downstream applications that analyze variability: First, sample-specific content that is missing from reference genomes is available for comparative analysis. Second, known topological features, such as those induced by structural variations or lateral gene transfers, can be directly identified in the graph and associated with various phenotypes (e.g., host disease) by examining the originating samples of each sequence. Finally, new and complex topological features of importance can be identified directly by examining the topology of the graph in light of such associations with phenotypes. Beyond the study of variation, we posit that almost every analysis could be performed and potentially improved by considering the pan-metagenomic graph. For example, by applying binning algorithms (22, 23), sequences in the graph can be assigned to their harboring microbes and taxonomically classified. Additionally, the graph itself can be used as a reference, using read-to-graph mapping methods (24). Finally, sequence coverage per sample can be used to calculate gene and taxon abundance estimates; it has been shown that such estimates are improved by consideration of shared genomic elements (2, 25), which are comprehensively available from the topology of the graph.

Our vision is to use our framework as a bedrock for an unbiased and systematic study of the pan-metagenome and its interactions with the host. We are developing methods that directly analyze this pan-metagenomic graph, doing away with the prevailing separation between the assembly and analysis stages. These methods have access to complete information about variability and the genomic topology encoded in our graph, which is typically unavailable with current analysis pipelines. In the coming years, we hope to leave behind the genome-centric model, and instead use high-resolution analyses of the pan-metagenome to accelerate our understanding of how genomic variability shapes the relation between the host and the microbiome.

ACKNOWLEDGMENTS

We thank members of the Korem lab and David Zeevi for useful discussions.

We acknowledge the support of the Program for Mathematical Genomics at Columbia University. T.K. is a CIFAR Azrieli Global Scholar in the Humans & the Microbiome Program.

REFERENCES

1. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244. <https://doi.org/10.1038/nature10571>.
2. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, Weinberger A, Fu J, Wijmenga C, Zhernakova A, Segal E. 2019. Structural variation in the gut microbiome associates with host health. *Nature* 568:43–48. <https://doi.org/10.1038/s41586-019-1065-y>.
3. Greenblum S, Carr R, Borenstein E. 2015. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160:583–594. <https://doi.org/10.1016/j.cell.2014.12.038>.

4. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. 2019. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe* 25:656–667.e8. <https://doi.org/10.1016/j.chom.2019.03.007>.
5. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013. Genomic variation landscape of the human gut microbiome. *Nature* 493:45–50. <https://doi.org/10.1038/nature11711>.
6. Kent AG, Vill AC, Shi Q, Satlin MJ, Brito IL. 2020. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat Commun* 11:4379. <https://doi.org/10.1038/s41467-020-18164-7>.
7. Sommer MOA, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325: 1128–1131. <https://doi.org/10.1126/science.1176950>.
8. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. 2001. Evidence for extensive resistance gene transfer among *Bacteroides* spp. and among *Bacteroides* and other genera in the human colon. *Appl Environ Microbiol* 67: 561–568. <https://doi.org/10.1128/AEM.67.2.561-568.2001>.
9. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568:505–510. <https://doi.org/10.1038/s41586-019-1058-x>.
10. Maguire F, Jia B, Gray K, Lau WYV, Beiko RG, Brinkman FSL. 2020. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb Genom* 6:mgen000436. <https://doi.org/10.1099/mgen.0.000436>.
11. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>.
12. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 40:e155. <https://doi.org/10.1093/nar/gks678>.
13. Li D, Liu C-M, Luo R, Sadakane K, 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. <https://doi.org/10.1093/bioinformatics/btv033>.
14. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27:824–834. <https://doi.org/10.1101/gr.213959.116>.
15. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829. <https://doi.org/10.1101/gr.074492.107>.
16. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
17. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11:2864–2868. <https://doi.org/10.1038/ismej.2017.126>.
18. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176:649–662.e20. <https://doi.org/10.1016/j.cell.2019.01.001>.
19. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 2019. A new genomic blueprint of the human gut microbiota. *Nature* 568:499–504. <https://doi.org/10.1038/s41586-019-0965-1>.
20. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* 27:140–153.e9. <https://doi.org/10.1016/j.chom.2019.10.022>.
21. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37:540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
22. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359. <https://doi.org/10.7717/peerj.7359>.
23. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32:605–607. <https://doi.org/10.1093/bioinformatics/btv638>.
24. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* 21:35. <https://doi.org/10.1186/s13059-020-1941-7>.
25. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje GB, Clement MJ, Crandall KA, Johnson WE. 2013. Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res* 23:1721–1729. <https://doi.org/10.1101/gr.150151.112>.