AMERICAN SOCIETY FOR MICROBIOLOGY mSystems®

COMMENTARY

Elucidating the Beta-Diversity of the Microbiome: from Global Alignment to Local Alignment

Xiaoquan Su[a,b]

[a]College of Computer Science and Technology, Qingdao University, Qingdao, China
[b]Single-Cell Center, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, China

ABSTRACT Quantitative comparison among microbiomes can link microbial beta-diversity to environmental features, thus enabling prediction of ecosystem properties or dissection of host-microbiome interaction. However, to compute beta-diversity, current methods mainly employ the entire community profiles of taxa or functions, which can miss the subtle differences caused by low-abundance community members that may play crucial roles in the properties of interest. In this work, I review the distance metrics and search engines that we developed to match microbiomes at a large scale based on whole-community-level similarities, as well as their limitations in tackling the microbiome changes caused by less abundant community features. Then I propose the concept of microbiome "local alignment," including an algorithm to measure microbiome similarity on specific fractions of biodiversity and an indexing strategy for rapidly fetching microbiome local-alignment matches from the data repository.

KEYWORDS microbiome, beta-diversity, distance metrics, search engine, local alignment

Beta-diversity analysis quantifies the similarity or distance between microbiome pairs; on the basis of beta-diversity analysis, we can link the overall taxonomic or functional diversity pattern to environmental features (1) and then predict the ecosystem properties or host healthy states (2–4). Here, I summarize the algorithms and tools that we have developed for analyzing and unitizing the whole-community-level (i.e., "global") similarities on large-scale microbiome data sets and deliver our perspective on the "local alignment" strategy that matches microbiomes by a specific subset of taxa that contribute to the properties of interest.

SIMILARITY MEASUREMENT FOR MICROBIOMES

An accurate and reliable similarity or distance metric among microbiomes is the basis for deducing the microbial beta-diversity. Statistical or geometry approaches like Bray-Curtis, Jaccard, and Jensen-Shannon divergence calculate such distances mainly by counting the overlapped components. However, omission of the inherent relationships among community members (e.g., operational taxonomic units for 16S rRNA amplicons or species for shotgun metagenomes) can lead to unexpected, erroneous beta-diversity patterns. To tackle this issue, we introduced the Meta-Storms scoring algorithm that parses the similarity of two microbiomes by considering the evolutionary hierarchy of microbes based on a weighted reference phylogeny tree (5). It not only improves the comprehensiveness of comparison by integrating additional biological contexts but also reduces the inaccuracy caused by the sparse distribution of microbes (e.g., microbiomes collected from distinct ecosystems may lack adequate common components for comparison) (6).

Citation Su X. 2021. Elucidating the beta-diversity of the microbiome: from global alignment to local alignment. mSystems 6:e00363-21. https://doi.org/10.1128/mSystems.00363-21.

Address correspondence to suxq@qdu.edu.cn.

Conflict of Interest Disclosures: X.S. 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

On the other hand, a phylogeny-based algorithm such as Meta-Storms requires all community members are mapped to definite leaf nodes in a reference tree; however, profiles inferred from metagenomic shotgun sequences always carry unidentified or unclassified annotations. To solve this problem, we then proposed the Dynamic Meta-Storms algorithm (7), by locating the unclassified species to the virtual nodes in the phylogeny tree via their higher-level taxonomy. Usually, the tree-like algorithm is well defined by a recursive posttraversal process of a binary tree. However, since the microbial phylogeny tree has been greatly expanded by newly sequenced and annotated species, the overall computing time for the pairwise distance matrix becomes unacceptable, especially for studies with thousands of samples. Hence, optimizations including nonrecursive transformation and memory recycling were performed in Meta-Storms and Dynamic Meta-Storms to improve the efficiency of computing and memory resource (8). Coupled with parallel computing on a multicore CPU (central processing unit) or a GPU (graphics processing unit), our implementations accomplished the pairwise comparison of 100,000 metagenomes within a few hours on a single desktop computer, enabling beta-diversity analysis on a much broader scale.

MICROBIOME SEARCH ENGINE ENABLES THE GLOBAL MATCH IN MICROBIOME DATA SPACE

Over the past years, the number of sequenced microbiomes has grown exponentially. While big data introduces a plethora of opportunities to uncover biological principles hidden under biodiversity surveys, new challenges have emerged, such as the extremely high data volume (9). One key demand and bottleneck has been relating newly sampled microbiomes to existing data. Thus, we developed a Microbiome Search Engine (MSE) for rapid search of query microbiomes against a database of microbiomes, on the whole-community level (10). Basically, with a given query community, MSE compares it against a data repository and returns top hits with highest Meta-Storms similarity in real time (e.g., <0.5 s per query in 1 million samples). This allows interpreting the property of the query based on meta-data of the matches. Moreover, by placing each individual sample under the context of the numerous microbiomes produced so far, MSE provides a bird's-eye view on the historical development of global microbiome surveying efforts. For example, tracking the 8-year dynamics of search-based microbiome novelty score (MNS) (which evaluates the overall compositional uniqueness of a microbiome compared to its top hits in a database) for more than 100,000 samples from various habitats, we were able to define the "search boundary effect" of human microbiomes (11). Specifically, the structural novelty of human microbiomes, but not environmental ones, is approaching saturation and likely bounded. More importantly, exploring the ability to quantitatively assess microbiome "novelty" or "uniqueness" via MNS, we introduced a search-based strategy for multiple disease detection and classification (12). In this method, MSE detects unhealthy samples via their outlier novelty versus a database of samples from healthy subjects and then identifies the specific disease type by comparing these to samples from patients. We showed that accuracy and efficiency of such MSE-based disease diagnosis outperform traditional machine learning approaches. These findings highlight the promise of microbiome big-data-based diagnosis as well as "data-driven" research strategies in microbiome science.

LOCAL ALIGNMENT FOR MICROBIOME FRACTIONS

Usually, beta-diversity is measured by end-to-end comparison of microbiome pairs (Fig. 1A) using distance metrics like Meta-Storms, UniFrac (13), Bray-Curtis, etc. The beta-diversity-based status identification and classification relies on an assumption that most members of the community, or at least the highly abundant members, are associated with the status of interest, e.g., samples in disease group exhibit a significant compositional distinction to healthy controls (e.g., permutational multivariate analysis of variance [PERMANOVA] or analysis of similarity [ANOSIM] test *P* value of <0.01 on pairwise distances). Although previous studies have shown such beta-diversity patterns exist in many diseases such as

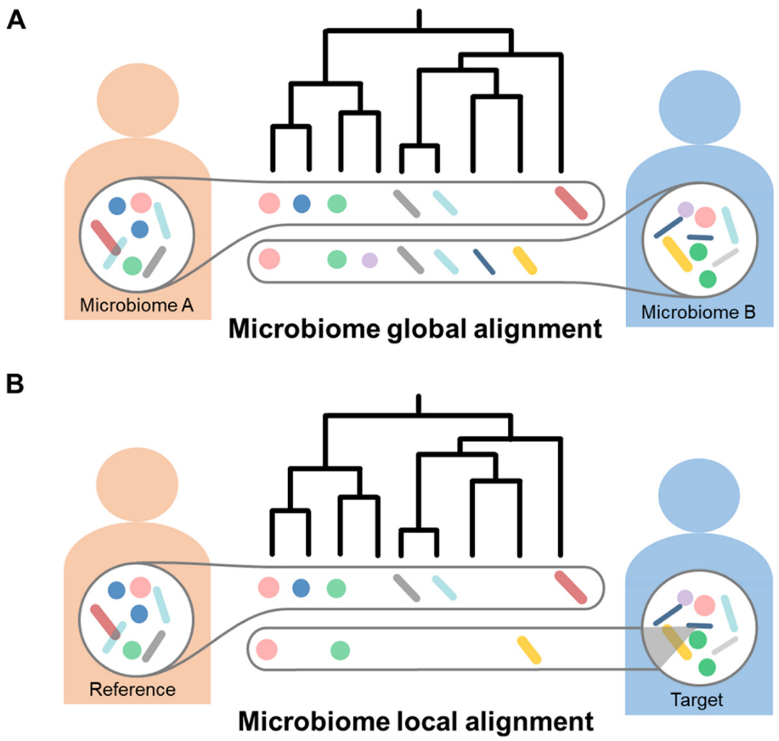

FIG 1 Two scenarios of microbiome comparison. (A) The end-to-end comparison of sample pairs employs whole-community-level information (i.e., "global alignment"). (B) The "local alignment" of microbiomes matches only a partial fraction of taxa that are of interest.

inflammatory bowel disease (14) and colorectal cancer (15), in other cases like type 1 diabetes (16) and autism spectrum disorder (17), only a small part of signature taxa play crucial roles that can be determined by statistical tests (18) or supervised machine learning (19) but are missed by the end-to-end comparison at the whole-community level. Thus, there is an intensive need to match only the "biomarker" fractions of interest (denoted as "target") against whole microbiomes (denoted as "reference"; Fig. 1B), just like a "local alignment" of amplified DNA fragments to the reference full-length 16S rRNA genes. Intuitively, such subcommunity-level similarity can be derived by extracting the identical features as the target from the reference and then compared it to the target. However, several issues should be appropriately covered in algorithm design and implementation. Since microbiome profiles are highly diverse and sparse across habitats (20) or cohorts (21), it is possible that a reference microbiome shares few exactly identical fractions with a target. Here, the similarity cannot be simply set as zero, and taxa with very close taxonomy or metabolic functions to the target or belong to the same guild (22) that work consistently and coherently with the target, can be considered "approximate members." Notably, contributions of such "approximate members" should be weighted by their phylogenetic or functional distances to the "exact members." On the other hand, however, once the "approximate members" are added for comparison, relative abundance of "exact members" will be diluted, leading to a reduction of similarity between reference and target. Therefore, for microbiome local alignment, selecting and extracting the fraction of community members from the reference microbiomes for the comparison to the target is of utmost importance.

INDEXING STRATEGY FOR FAST FETCH OF LOCAL-ALIGNMENT HITS

Once the microbiome "local alignment" algorithm is clearly defined, suspected unhealthy microbiomes can be detected from a repository by matching with specific disease biomarkers. An exhaustive screening that compares the target fractions to all samples is a straightforward way, but it is time-consuming when the database is huge.

Currently, there are two types of indexing strategies available for accelerating the microbiome search, (i) a static partitions index that groups database into subcategories sorted by structural features, e.g., Microbiome Search Engine v1.0 (5) or Meta-Prism (23); (ii) a dynamic index based on the dimension reduction of microbial profiles employed by Microbiome Search Engine 2 (10). Both of the approaches depend on the preprocessing of the entire collection of reference samples in the database construction step in order to rapidly fetch the candidate hits in the subsequent query step. Nevertheless, as the "local alignment" only takes partial community from the reference, and the range of community members relies on the specific query target (e.g., biomarkers for diseases), unified and universal indices designed for end-to-end match are not suitable for the "local alignment" scenario. A potential indexing solution to promote the speed of microbiome local-alignment can learn from the FM-index of Bowtie 2 (24) or the USEARCH algorithm (25) that were originally designed for nucleotide sequence mapping in which the target community fraction serves as a short query DNA read and the microbiomes are treated as the reference long genome sequences.

CONCLUSION

Beta-diversity is a fundamental property of microbiomes. Highly efficient microbiome comparison, not just at the "global" level but at the "local" level, can elucidate microbial beta-diversity with higher precision and flexibility, thus contributing to in-depth comprehension and efficient utilization of microbiomes.

ACKNOWLEDGMENTS

I acknowledge the support of grants 31771463 and 32070086 from the National Natural Science Foundation of China.

I thank Jian Xu for helpful discussions.

REFERENCES

1. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410–422. https://doi.org/10.1038/s41579-018-0029-9.
2. Teng F, Yang F, Huang S, Bo C, Xu ZZ, Amir A, Knight R, Ling J, Xu J. 2015. Prediction of early childhood caries via spatial-temporal variations of oral microbiota. Cell Host Microbe 18:296–306. https://doi.org/10.1016/j.chom.2015.08.005.
3. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567–574. https://doi.org/10.1038/s41586-020-2095-1.
4. Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, Gasbarrini A, Tortora G. 2020. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol 17:635–648. https://doi.org/10.1038/s41575-020-0327-3.
5. Su X, Xu J, Ning K. 2012. Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics 28:2493–2501. https://doi.org/10.1093/bioinformatics/bts470.
6. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.
7. Jing G, Zhang Y, Yang M, Liu L, Xu J, Su X. 2020. Dynamic Meta-Storms enables comprehensive taxonomic and phylogenetic comparison of shotgun metagenomes at the species level. Bioinformatics 36:2308–2310. https://doi.org/10.1093/bioinformatics/btz910.
8. Su X, Wang X, Jing G, Ning K. 2014. GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU. Bioinformatics 30:1031–1033. https://doi.org/10.1093/bioinformatics/btt736.
9. Su X, Jing G, Zhang Y, Wu S. 2020. Method development for cross-study microbiome data mining: challenges and opportunities. Comput Struct Biotechnol J 18:2075–2080. https://doi.org/10.1016/j.csbj.2020.07.020.
10. Jing G, Liu L, Wang Z, Zhang Y, Qian L, Gao C, Zhang M, Li M, Zhang Z, Liu X, Xu J, Su X. 2021. Microbiome Search Engine 2: a platform for taxonomic and functional search of global microbiomes on the whole-microbiome level. mSystems 6:e00943-20. https://doi.org/10.1128/mSystems.00943-20.
11. Su X, Jing G, McDonald D, Wang H, Wang Z, Gonzalez A, Sun Z, Huang S, Navas J, Knight R, Xu J. 2018. Identifying and predicting novelty in microbiome studies. mBio 9:e02099-18. https://doi.org/10.1128/mBio.02099-18.
12. Su X, Jing G, Sun Z, Liu L, Xu Z, McDonald D, Wang Z, Wang H, Gonzalez A, Zhang Y, Huang S, Huttley G, Knight R, Xu J. 2020. Multiple-disease detection and classification across cohorts via microbiome search. mSystems 5:e00150-20. https://doi.org/10.1128/mSystems.00150-20.
13. McDonald D, Vázquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, Gonzalez A, Knight R. 2018. Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat Methods 15:847–848. https://doi.org/10.1038/s41592-018-0187-8.
14. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, McClure EE, Dunklebarger MF, Knight R, Jansson JK. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2:17004. https://doi.org/10.1038/nmicrobiol.2017.4.
15. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25:679–689. https://doi.org/10.1038/s41591-019-0406-6.
16. Alkanani AK, Hara N, Gottlieb PA, Ir D, Robertson CE, Wagner BD, Frank DN, Zipris D. 2015. Alterations in intestinal microbiota correlate with

susceptibility to type 1 diabetes. Diabetes 64:3510–3520. https://doi.org/10.2337/db14-1847.

17. Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, Litcher-Kelly L, Gadow KD, Gathungu G, Robertson CE, Ir D, Frank DN, Li E. 2015. Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the Simons Simplex Collection. PLoS One 10: e0137725. https://doi.org/10.1371/journal.pone.0137725.
18. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60. https://doi.org/10.1186/gb-2011-12-6-r60.
19. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977.
20. Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, Schulze-Lefert P. 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host Microbe 17:603–616. https://doi.org/10.1016/j.chom.2015.04.009.
21. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 27:626–638. https://doi.org/10.1101/gr.216242.116.
22. Wu G, Zhao N, Zhang C, Lam YY, Zhao L. 2021. Guild-based analysis for understanding gut microbiome in human health and diseases. Genome Med 13:22. https://doi.org/10.1186/s13073-021-00840-y.
23. Zhu M, Kang K, Ning K. 2021. Meta-Prism: ultra-fast and highly accurate microbial community structure search utilizing dual indexing and parallel computation. Brief Bioinform 22:557−567. https://doi.org/10.1093/bib/bbaa009.
24. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
25. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.