JCI The Journal of Clinical Investigation

Guild-based approach for mitigating information loss and distortion issues in microbiome analysis

Liping Zhao, … , Guojun Wu, Naisi Zhao

J Clin Invest. 2024;134(17):e185395. https://doi.org/10.1172/JCI185395.

Viewpoint

Introduction Microbiome research holds the promise of elucidating new mechanisms of disease development and developing innovative approaches for disease prevention and treatment. However, a substantial challenge that stymies this potential is the often overlooked issue of information loss and distortion during data analysis. Such analytical shortcomings are a primary contributor to inconsistencies between studies. For example, conflicting findings often emerge, such as divergent taxa linked to the same diseases in separate studies (1, 2). Likewise, the same taxon may have contrasting associations with identical diseases across different investigations (2, 3). This issue is exemplified by the phylum Firmicutes, which has been linked to both an increase and a decrease in prevalence of type 2 diabetes across different studies (4, 5). Similar inconsistencies are found at the genus level; conflicting reports exist regarding the role of Collinsella in autism spectrum disorder. While studies by Strati et al. (6) and Chamtouri et al. (7) found a positive association (detrimental effects) between Collinsella and autism spectrum disorder, other researchers showed a reduction of Collinsella (8, 9). Such inconsistencies are common in microbiome studies on various diseases (2). The root of these inconsistencies often lies in the insufficient recognition of the profound genetic and functional diversity present at the strain level within a single bacterial species. The average nucleotide identity within a bacterial species […]

Guild-based approach for mitigating information loss and distortion issues in microbiome analysis

Liping Zhao,[1,2] Guojun Wu,[1] and Naisi Zhao[3]

[1]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences and Center for Microbiome, Nutrition, and Health, New Jersey Institute for Food, Nutrition, and Health, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA. [2]State Key Laboratory of Microbial Metabolism and Ministry of Education Key Laboratory of Systems Biomedicine, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [3]Department of Public Health and Community Medicine, School of Medicine, Tufts University, Boston, Massachusetts, USA.

Introduction

Microbiome research holds the promise of elucidating new mechanisms of disease development and developing innovative approaches for disease prevention and treatment. However, a substantial challenge that stymies this potential is the often overlooked issue of information loss and distortion during data analysis. Such analytical shortcomings are a primary contributor to inconsistencies between studies. For example, conflicting findings often emerge, such as divergent taxa linked to the same diseases in separate studies (1, 2). Likewise, the same taxon may have contrasting associations with identical diseases across different investigations (2, 3). This issue is exemplified by the phylum Firmicutes, which has been linked to both an increase and a decrease in prevalence of type 2 diabetes across different studies (4, 5). Similar inconsistencies are found at the genus level; conflicting reports exist regarding the role of Collinsella in autism spectrum disorder. While studies by Strati et al. (6) and Chamtouri et al. (7) found a positive association (detrimental effects) between Collinsella and autism spectrum disorder, other researchers showed a reduction of Collinsella (8, 9). Such inconsistencies are common in microbiome studies on various diseases (2).

The root of these inconsistencies often lies in the insufficient recognition of the profound genetic and functional diversity present at the strain level within a single bacterial species. The average nucleotide identity within a bacterial species can vary by 4% to 5% (10), a striking contrast when compared with the approximate 1% genomic difference between humans and chimpanzees (11). This heterogeneity within a bacterial species demands our attention, for it holds the key to comprehending the delicate balance within microbiomes. Despite advances in technology that allow for analyses at finer granularities, such as amplicon sequence variant (ASV) (12) and metagenome-assembled genome (MAG) (13), conventional data analysis methodologies in microbiome research often fail to account for this strain-level variation. This oversight leads to a cascade of information loss and distortion, ultimately impeding our comprehension of the intricate connections between the microbiome and human health.

In this Viewpoint article, we will dissect the limitations that hinder our strain-level understanding, delve into tools for evaluating information loss and distortion, and advocate for a genome-centric and guild-based approach to mitigating these issues (Figure 1). Such a paradigm shift is not only about refining technical approaches, but is also about adopting a new perspective that can enhance the integrity and applicability of microbiome research.

Evaluating information loss and distortion

Microbiome analysis generates extensive datasets composed of unique sequences, such as ASVs or MAGs, each representing unique types of microbes. These datasets encapsulate a wealth of information about microbial diversity and functionality within microbiome samples. Nevertheless, the high dimensionality and sparsity of these datasets present significant challenges. With variables outnumbering samples, a phenomenon known as the "curse of dimensionality" emerges, complicating the identification of authentic health-related microbial signatures (14). Thus, reducing the dimensionality and sparsity of the original microbiome datasets, collectively called data reduction, is imperative for microbiome analysis. However, information loss and distortion can occur in current data reduction practices in mainstream microbiome analysis.

Information loss. Information loss on novel or understudied microbes and their functions can occur in database-dependent analysis of microbiome datasets. The primary step in conventional microbiome data analysis involves taxonomic assignment and functional annotation, heavily relying on reference databases such as SILVA (15) (https://www.arb-silva.de/) or KEGG (16) (https://www.genome.jp/kegg/). When unclassified or unannotated sequences are excluded from downstream analysis, information on the diversity and function of novel or understudied microbes they represent will be ignored in any further analysis. In practice, it's common for 10%–40% of ASVs to remain unclassifiable at the genus level, and up to 50% of genes may lack functional annotations (17). This exclusion can result in a substantial portion of data being disregarded, thus potentially skewing the representation of microbial communities and functions.

Information distortion. Information distortion, on the other hand, happens when the process of reducing dataset complexity introduces biases. For instance, lumping ASVs by genus or genes by pathways (14) can conceal the nuances of strain-level variation. Strains within the same taxon may exhibit different or opposing correlations with the same disease or intervention. Similarly, the same critical pathway gene, such as the *but* gene for butyrate production, may be harbored by two competing bacterial strains, masking

Conflict of interest: LZ is a cofounder of Notitia Biotechnologies Co.

Reference information: *J Clin Invest*. 2024;134(17):e185395. https://doi.org/10.1172/JCI185395.

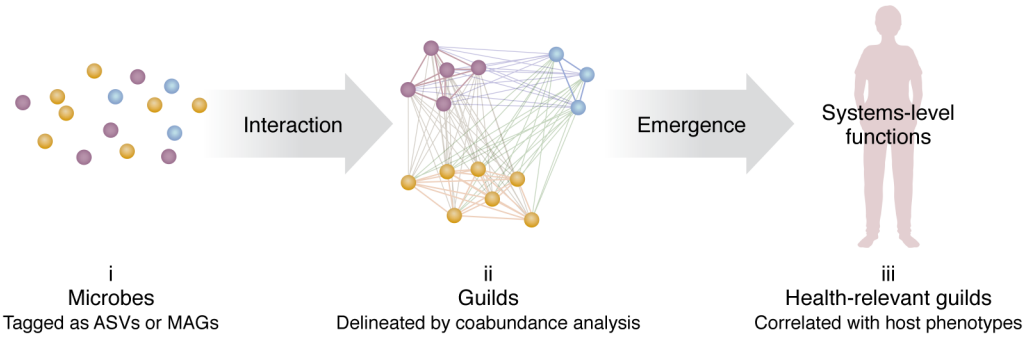

Figure 1. Guild-based analysis of microbiome datasets. A guild consists of microbes with diverse taxonomic backgrounds, but thriving or declining together, showing coabundance behavior. (i) Initial tagging and identification of individual microbial entities using ASVs or MAGs, each assigned a UUID for precise tracking. (ii) Analysis of interactions among microbial entities to identify patterns of coexistence and influence, revealing the foundational relationships within the microbiome. Clustering of microbes into guilds based on coabundance analysis, where members share ecological niches and exhibit similar abundance patterns across different conditions or samples. (iii) Emergence of complex microbial behaviors and functions from the interactions within and between guilds, highlighting the collective capabilities of the microbiome. Integration of guild activities into broader systems-level functions that affect host physiology and health, encapsulating the holistic effect of microbial interactions.

the true abundance change of this gene in microbiome datasets (18). Failure to account for these strain-level variations during dimensionality reduction can lead to information distortion, resulting in inconsistencies across microbiome studies and hindering the establishment of clear associations between microbiome features and diseases.

A model for evaluating information loss and distortion. To address these challenges, we propose employing β diversity matrices of all ASVs or MAGs as a benchmark for the entire information content of the original datasets. We then advocate for the combined use of Procrustes analysis (19) and the Mantel test (20) to evaluate information loss and distortion, which may happen after each attempt at data reduction. These methods can compare and assess the similarity or dissimilarity between multivariate datasets. A close match between the β diversity matrices before and after data reduction indicates successful preservation of original dataset characteristics. At the same time, a significant difference may signal a loss or misrepresentation of information. For example, a new β diversity matrix based on genus-level variables should be created when analyzing ASV datasets at the genus level. This new matrix needs to be compared with the original one at the ASV level using Procrustes analysis and the Mantel test. If the matrices show congruence, it suggests minimal information loss and distortion, indicating that the reduced dataset accurately represents the original. Conversely, the pronounced disparity between the matrices may indicate potential information misrepresentation. In addition, when comparing different data reduction methods, the combined use of Procrustes analysis and the Mantel test can help determine which method better preserves information from the original datasets.

These methods ensure that dimensionality reduction maintains data integrity. By preserving the essence of information in the original datasets while reducing complexity, researchers can generate more reproducible and consistent results for microbiome biomarker discovery.

Mitigating information loss and distortion

We advocate for a guild-based analytical strategy to confront the pervasive issues of information loss and distortion in microbiome analysis (17). This innovative approach transcends the confines of traditional methods, offering a precise and ecologically sound representation of microbial communities. The guild-based approach is supported by three key pillars.

Genome-specific analysis. Microbial cells and viral particles are the fundamental units of change at the core of the gut ecosystem. A genome-specific approach leverages genomic data as molecular tags to track and catalog the entire microbial constituency. Advancements in sequencing technologies are bringing us closer to a future where comprehensive genomic mapping of microbiomes becomes feasible and cost-effective. Until then, ASVs or MAGs with a 1% average nucleotide identity (ANI) difference are proxies for such detail, allowing near strain-level resolution.

Database-independent inclusivity. To curtail information loss inherent in database-dependent analyses, we can implement a system of universal unique identifiers (UUIDs) for each MAG or ASV, streamlining tracking across samples and studies. The generation of UUIDs is solely based on the sequence identity between MAGs or ASVs. New UUIDs will be assigned if the novel MAGs and ASVs are not found in existing studies. With such a UUID system in place, taxonomic assignment or functional annotation will not be the primary step in microbiome analysis. Thus, novel microbes will not be excluded from downstream analysis. This reference-free approach ensures that our analysis remains unbiased toward known species, enabling the discovery of previously unidentified or understudied microbes. By embracing the unknown, we achieve a more inclusive and comprehensive view of the microbiome, reducing information loss related to database limitations.

Interaction-focused aggregation. In the intricate web of the gut ecosystem, microbes do not exist in isolation, but rather form synergistic collectives known as guilds. Members in the same guild cooperate, thrive, or decline together, showing coabundance behavior. Different guilds may cooperate or compete to form the whole ecosystem network. We introduce guild-level categorization as the primary method for dimensionality reduction

in microbiome analysis (17), focusing on functional groupings within ecosystems, which consider the complex interactions between microbes. Members in these guilds can be clustered together based on their coabundance behavior, irrespective of their taxonomic background. This perspective considers the ecological interactions and cooperative relationships among microbes, providing insights into how groups work together to influence microbiome stability and function. This approach ensures that valuable functional insights are not obscured by taxonomic lumping, minimizing the information distortion.

The guild-based approach ushers in a more objective, holistic, and functionally oriented understanding of microbial communities and their impact on human health. This framework has revealed bacterial guilds' potential role in disease phenotype development, such as obesity in Prader-Willi syndrome (21), and uncovered microbial guilds alleviating type 2 diabetes when fostered by dietary fibers (18). Our recent study further demonstrates the power of the guild-based analytical approach, structured around the three methodological pillars: genome specificity, database independence, and interaction-focused aggregation. By focusing on stably correlated genomes, we identified a core microbiome characterized by two competing guilds, one beneficial, the other detrimental. This core structure persisted despite the diverse confounding factors inherent in microbiome datasets spanning various studies with wide variations of interventions, diseases, geographic locations, ethnicities, and sequencing protocols. This finding underscores the robustness of the guild-based approach in capturing fundamental microbiome patterns that are crucial for understanding human health (22). By implementing this approach, we can effectively mitigate information loss and distortion, thereby enhancing the robustness and reproducibility of microbiome research and improving the validity of our findings across studies.

Conclusion

As microbiome research advances, the imperative to overcome information loss and distortion becomes increasingly critical. The genome-centric and guild-based methodologies represent our commitment to this cause. We aspire to mitigate these challenges by adopting genome-centric and guild-based analysis. In this quest for precision and comprehensiveness, we extend an invitation to the global research community. We call upon the global research community to join in refining these approaches, thus fortifying the integrity of microbiome research and catalyzing breakthroughs in disease prevention, diagnosis, and treatment, with far-reaching implications spanning science, medicine, and beyond.

Address correspondence to: Liping Zhao, Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, Lipman Hall–Room 326, 76 Lipman Drive, New Brunswick, New Jersey 08901-8525, USA. Email: liping.zhao@rutgers.edu.

1. Walker AW, Hoyles L. Human microbiome myths and misconceptions. *Nat Microbiol*. 2023;8(8):1392–1396.
2. Dai D, et al. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res*. 2022;50(d1):D777–D784.
3. Yao G, et al. MicroPhenoDB associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. *Genomics Proteomics Bioinformatics*. 2020;18(6):760–772.
4. Ahmad A, et al. Analysis of gut microbiota of obese individuals with type 2 diabetes and healthy individuals. *PLoS One*. 2019;14(12):e0226372.
5. Larsen N, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One*. 2010;5(2):e9085.
6. Strati F, et al. New evidences on the altered gut microbiota in autism spectrum disorders. *Microbiome*. 2017;5(1):24.
7. Chamtouri M, et al. Age and severity-dependent gut microbiota alterations in Tunisian children with autism spectrum disorder. *Sci Rep*. 2023;13(1):18218.
8. De Angelis M, et al. Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified. *PLoS One*. 2013;8(10):e76993.
9. Peralta-Marzal LN, et al. A robust microbiome signature for autism spectrum disorder across different studies using machine learning. *Sci Rep*. 2024;14(1):814.
10. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*. 2009;106(45):19126–19131.
11. King M-C, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*. 1975;188(4184):107–116.
12. Callahan BJ, et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581–583.
13. Bowers RM, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*. 2017;35(8):725–731.
14. Armstrong G, et al. Applications and comparison of dimensionality reduction methods for microbiome data. *Front Bioinform*. 2022;2:821861.
15. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(database issue):D590–D596.
16. Kanehisa M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44(d1):D457–D462.
17. Wu G, et al. Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Med*. 2021;13(1):22.
18. Zhao L, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science*. 2018;359(6380):1151–1156.
19. Peres-Neto PR, Jackson DA. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*. 2001;129(2):169–178.
20. Legendre P, Legendre L, eds. *Numerical Ecology*. Elsevier; 2012.
21. Zhang CH, et al. Dietary modulation ofh gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*. 2015;2(8):968–984.
22. Wu G, et al. A core microbiome signature as an indicator of health. *Cell*. In press.