bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Quantification of human gut microbiota variability using replicate sampling and spike-in sequencing

Brian W. Ji[1,2*], Ravi U. Sheth[1,2*], Purushottam D. Dixit[1], Harris H. Wang[1,3†] & Dennis Vitkup[1,4†]

[1]Department of Systems Biology, Columbia University, New York, NY, USA.
[2]Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University, New York, NY, USA.
[3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA
[4]Department of Biomedical Informatics, Columbia University, New York, NY, USA.

*These authors contributed equally to this work.
[†]To whom correspondence should be addressed.

Abstract

16S rRNA amplicon sequencing has enabled detailed investigation of the spatiotemporal dynamics of the human gut microbiome and their corresponding alterations during disease[1–4]. However, the contributions of temporal changes, spatial sampling location and technical sources of variability in gut microbiome studies remain poorly understood. Commonly used sequencing approaches are further limited by compositional biases due to relative abundance measurements[5–7]. Here, we combine mathematical modeling with an experimental workflow based on replicate sampling and spike-in sequencing to separately quantify the major sources of gut microbiota variability measured in absolute abundances. We apply this framework to the healthy human gut microbiome and find substantial and distinct contributions to measured abundance variability associated with time, spatial sampling location and technical noise. Notably, our approach allows us to identify a critical abundance threshold (~0.01% in average taxa relative abundance) above which variability primarily results from temporal changes and below which technical noise predominates. Furthermore, we find a large contribution (~20%) to measured microbiota variability resulting from different spatial sampling locations. Across all taxa, we observe that overall patterns of temporal and spatial variability in the human gut microbiome follow closely those in other diverse ecosystems[8,9], but we also identify specific taxa whose behavior are largely associated with either underlying temporal or spatial sources. Finally, we use our approach to show that temporal factors can largely explain the significant changes in total bacterial densities observed in the gut and the abundance correlations of individual bacterial taxa. Collectively, our results highlight important pitfalls of current fecal profiling practices and provide a general framework to facilitate future quantitative ecological analysis of the human gut microbiome and other complex microbial communities.

Main Text

Metagenomic sequencing is now widely utilized to survey microbial diversity and taxa abundances of fecal microbiota[10–14]. In a commonly used approach, genomic DNA is extracted from samples and the 16S rRNA region is then amplified and sequenced. Resulting reads are clustered into operational taxonomic units (OTUs), thus yielding relative abundances of individual bacterial taxa within each sample. Abundances obtained in this way have been used to investigate disease pathophysiology[15–17], microbe-microbe interactions[18–20] and microbiome dynamics[1,21–23].

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Despite their widespread use, current 16S rRNA sequencing approaches have important limitations. First, there are no principled methods that can separate temporal sources of gut microbiota variability from that of spatial sampling location[20,24] and confounding technical sources[10,25]. Second, OTU measurements are typically reported in relative abundances, which may mask underlying absolute abundance changes of bacteria[5–7]. With the exponentially increasing scale of microbiome studies[2,26,27] and emerging microbiota diagnostic paradigms[28,29], a quantitative understanding of the primary sources of microbial variability will be critical for the proper analysis and interpretation of these data.

To address these challenges, we developed a custom fecal sampling approach and complementary mathematical model to separately quantify sources of bacterial abundance variability in the gut, as well as a spike-in procedure during sample processing[30] to estimate total bacterial loads (Fig. 1a, Box 1, Supp. Note, Methods). In our Decomposition of Variance using Replicate Sampling (DIVERS) approach, fecal samples are collected from two independent locations of the same stool specimen on each day of the time series study. One of these samples is then further processed in duplicate. Interestingly, this sampling scheme and underlying model mirror the classic dual reporter approach used previously to separate intrinsic and extrinsic sources of noise in gene expression profiles[31–33]. The developed approach relies on the fact that OTUs that experience genuine temporal fluctuations should exhibit high covariance of spatial replicates across different time points, with spatial sampling variability and technical noise acting to diminish this correlation. Similarly, spatial sampling variability should lead to high covariance of technical replicates across different sampling locations after controlling for temporal correlations.

Box 1: Decomposition of bacterial abundance variance

Let X denote the total bacterial density in a collected sample or the abundance of an individual OTU. The variance of X can be written as a sum of three components associated with temporal, spatial and technical factors contributing to changes in X across samples:

$$Var(X) = \underbrace{Var_T E_{S|T} E(X|S,T)}_{\text{Temporal}} + \underbrace{E_T Var_{S|T} E(X|S,T)}_{\text{Spatial}} + \underbrace{E_T E_{S|T} Var(X|S,T)}_{\text{Technical}} \quad (1)$$

Here, S and T are space and time-associated random variables capturing the respective spatial and temporal processes influencing the abundance of X from sample to sample. Following the notation in Figure 1, each of these terms are estimated as follows (see Supplementary Note for full derivations):

$$\underbrace{Var_T E_{S|T} E(X|S,T)}_{\text{Temporal}} = Cov(X,Z) \quad (2)$$

$$\underbrace{E_T Var_{S|T} E(X|S,T)}_{\text{Spatial}} = Cov(X-Z,Y) \quad (3)$$

$$\underbrace{E_T E_{S|T} Var(X|S,T)}_{\text{Technical}} = \frac{1}{2} Var(X-Y) \quad (4)$$

where X, Z and Y, Z denote spatial replicate pairs of either total bacterial density measurements or individual OTU abundances, and X, Y denote technical replicates.

We first characterized total baseline bacterial abundance variation in the human gut microbiome by applying DIVERS to a healthy male individual tracked daily over the course of three weeks. Consistent with previous findings[5], we found that total bacterial densities (measured in arbitrary

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

units) fluctuate significantly across different days and sampling locations (coefficient of variation ~0.5) (Fig. 1b). Notably, the observed variability resulted primarily from daily temporal changes in total load (Fig. 1b). Both spatial sampling variation and technical noise made significantly smaller contributions, demonstrating that at the macroscopic scale of our fecal sample collection, total bacterial loads on each day remain relatively constant across different spatial locations.

Utilizing these measurements of total bacterial density in each sample, we transformed relative OTU abundances into absolute abundance units and used our model to decompose the abundance variance of individual OTUs (Methods, Supp. Note). We reasoned that bacteria with low abundances would likely be significantly affected by technical error whereas variations reflecting true biological phenomena would be more pronounced for higher abundance OTUs. Indeed, variance profiles exhibited two distinct regimes as a function of average OTU abundance across samples, with a transition occurring at ~0.01% in relative abundances (Fig. 1c, Supp. Fig. 1). Fluctuations of OTUs below this cutoff largely reflected technical variability consistent with Poissonian sampling noise (Supp. Fig. 2b, Supp. Fig 3a,c). In contrast, the variability of higher abundance OTUs could be primarily explained by temporal changes, suggesting that OTU abundance fluctuations in the gut largely reflect underlying ecological dynamics, likely driven by environmental and dietary changes[1,34] and interspecies interactions[23,35] (Fig. 1c, Supp. Fig 2a). Interestingly, spatial sampling location also contributed a surprisingly large fraction of total variability (mean ~20% over abundant OTUs), demonstrating significant spatial heterogeneity of fecal samples despite the relatively large mass of fecal matter processed per sample (mean ~45 mg). As a validation for our variance decomposition model, we conducted a set of controlled experiments that specifically eliminated either temporal or spatial variability from our collected fecal samples. First, to remove any temporal variability from data, we applied our DIVERS protocol to a single stool specimen and obtained fecal samples from ten independent spatial locations, thus effectively simulating five consecutive days of sampling. Second, to remove spatial variability, we carried out eight consecutive days of replicate sampling with spatial replicates homogenized on each day before sequencing (Methods). Importantly, the model correctly predicted no temporal or spatial contributions to OTU abundance variability when the underlying signals were absent from data (Supp. Fig 4).

The ability to extract temporal and spatial variances from individual OTUs enables quantitative investigation of ecological relationships in the gut microbiome. Interestingly, temporal variances predicted from our model scaled strongly with average OTU abundance, following closely a relationship described by Taylor's power law in ecology[8,9] (Fig. 2a, power law exponent b = 1.87, R = 0.97). In addition, our replicate sampling approach allowed us to investigate the relationship between mean OTU abundance and spatial variance, which was also well-described by a power law (Fig. 2a, b = 1.70, R = 0.96). This suggests that in contrast to the null model of randomly distributed abundances (b = 1), bacterial species in the gut microbiome show significant spatial aggregation[20]. We note that as the model-estimated spatial variances in our analysis reflect an average over time (Supp. Note), they differ slightly from scaling observed from multiple spatial replicates obtained from a single day (Supp. Fig. 3b).

Differences in temporal and spatial abundance variance may also provide insight into the ecological behavior of phylogenetically diverse bacteria. When grouping abundant OTUs by the three most prevalent phyla in the gut, we found variance profiles on average to be similar across phyla, with temporal contributions comprising the majority of abundance variance (Supp. Fig. 5). This indicates that overall, there exists a consistent hierarchy to sources of variability in the gut. Despite this similar behavior on average across phyla, we used our model to identify specific taxa with particularly high spatial or temporal contributions (>70%) to total variance (Fig. 2c, Table S3, Methods). The time series of identified strains show divergent behavior of different individual

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bacteria over space and time, highlighting that within-sample spatial heterogeneity may confound analysis of longitudinal gut microbiome studies, at least for certain OTUs (Fig. 2d,e).

Changes in individual OTU abundances are often a result of the collective behavior of multiple different bacterial species[18,36]. We therefore used our mathematical framework to decompose the abundance covariance of all OTU pairs. (Box 2, Supp. Note, Methods). Analogous to the variance decomposition model, the total abundance covariance for any pair of OTUs may result from temporal sources driving changes in overall abundances on each day, spatial factors that cause different bacteria to co-occur or exclude one another across different sampling locations, and technical noise that may lead to spuriously correlated behavior across sequencing replicates. After normalizing covariance contributions by individual OTU standard deviations to obtain correlation contributions (Box 2), we found that the majority of pairwise abundance correlations could be attributed to temporal sources, with relatively minor contributions from sampling location and technical factors (Fig. 3a,b, Supp. Fig. 6) We also found, as reported previously[5], that total correlations based on absolute abundance measurements were larger on average than those calculated from relative abundances. This effect is likely due to the variance in total load across samples, which leads to an additional positive contribution to relative abundance correlations (Supp. Fig. 7, Supp. Note).

Box 2: Decomposition of bacterial abundance covariance

As with variances, the covariance between the abundances of any two OTUs i and j, denoted X_i and X_j, can also be written as a sum of temporal, spatial and technical contributions:

$$Cov(X_i, X_j) = \underbrace{Cov_T\left(E(X_i|T), E(X_j|T)\right)}_{\text{Temporal}} + \underbrace{E_T Cov_{S|T}\left(E(X_i|S,T), E(X_j|S,T)\right)}_{\text{Spatial}} + \underbrace{E_T E_{S|T} Cov(X_i|S,T, X_j|S,T)}_{\text{Technical}} \quad (5)$$

Each of these terms may also be estimated using the protocol described in Figure 1 as follows:

$$\underbrace{Cov_T\left(E(X_i|T), E(X_j|T)\right)}_{\text{Temporal}} = Cov(X_i, Z_j) \quad (6)$$

$$\underbrace{E_T Cov_{S|T}\left(E(X_i|S,T), E(X_j|S,T)\right)}_{\text{Spatial}} = Cov(X_i - Z_i, Y_j) \quad (7)$$

$$\underbrace{E_T E_{S|T} Cov(X_i|S,T, X_j|S,T)}_{\text{Technical}} = \frac{1}{2} Cov(X_i - Y_i, X_j - Y_j) \quad (8)$$

where X_i,Z_i and Y_i,Z_i denote spatial replicate measurements of the abundance of OTU i and X_i,Y_i denote technical replicates. To obtain temporal, spatial and technical correlations shown in Figure 3, we simply normalize each covariance contribution by the respective standard deviations of individual OTUs:

$$Cor(X_i, X_j) = \underbrace{\frac{Cov_T(E(X_i|T), E(X_j|T)}{\sigma_{X_i}\sigma_{X_j}}}_{\text{Temporal}} + \underbrace{\frac{E_T Cov_{S|T}\left(E(X_i|S,T), E(X_j|S,T)\right)}{\sigma_{X_i}\sigma_{X_j}}}_{\text{Spatial}} + \underbrace{\frac{E_T E_{S|T} Cov(X_i|S,T, X_j|S,T)}{\sigma_{X_i}\sigma_{X_j}}}_{\text{Technical}} \quad (9)$$

Full derivations for all equations are provided in the Supplementary Note.

Co-occurrence patterns across space and time may also reveal underlying ecological interactions of phylogenetically distinct taxa[7,19,37]. We therefore examined correlations of OTU abundances

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stratified across the three most prevalent phyla in gut by analyzing pairs of OTUs belonging to the same phyla (Fig. 3c). While the overall patterns of correlations were preserved within each of the phyla, OTUs belonging to the *Bacteroidetes* exhibited significantly larger temporal and total correlations among themselves as compared to other phyla ($p < 1e\text{-}10$, Wilcoxon rank sum test). This result remained significant even after controlling for 16S rRNA sequence dissimilarity (Fig. 3c, Supp. Fig. 8). Such correlated abundance fluctuations of *Bacteroidetes* may result from environmental factors such as dietary polysaccharide availability on each day[38,39] or interactions involving reciprocal cross-feeding observed in these taxa[40,41].

Quantification of the sources of variability in bacterial abundance measurements is important for understanding the underlying biological processes in the human gut microbiome. Here, we have developed and validated DIVERS, a novel data collection and modeling framework based on replicate sampling and spike-in sequencing. The developed framework allows the separation of temporal, spatial and technical sources of gut microbiota variability measured in absolute abundances. Importantly, our method corrects for compositional artifacts associated with relative abundance measurements and enables the identification of study and context-specific criteria for data inclusion and analysis. Furthermore, our method makes it possible to infer the underlying sources of variability of individual bacteria that are masked with current sequencing approaches.

More broadly, our study provides a quantitative framework for understanding general ecological relationships in diverse microbial communities. While our results suggest that human gut microbial ecology is highly dynamic over time, it will be interesting to compare results from the gut microbiome to other host-associated and non-host associated communities. It will also be interesting to determine how the variability associated with sampling location changes as the spatial resolution of sample collection is systematically altered. The flexibility of DIVERS should facilitate deep exploration of the primary determinants of microbial variation and co-occurrence patterns across highly diverse ecosystems.

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).
2. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550,** 61–66 (2017).
3. Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* **534,** 259–262 (2016).
4. Thaiss, C. A. *et al.* Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* **159,** 514–529 (2014).
5. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551,** 507–511 (2017).
6. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8,** 1–11 (2012).
7. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6,** 1–20 (2017).
8. Taylor, L. R. Aggregation, Variance and the Mean. *Nature* **189,** 732–735 (1961).
9. Taylor, L. R. & Woiwod, I. P. Temporal Stability as a Density-Dependent Species Characteristic. *J. Anim. Ecol.* **49,** 209–224 (1980).
10. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34,** 942–949 (2016).
11. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science (80-.).* **341,** 1237439 (2013).
12. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical innovations for high-throughput amplicon sequencing. *Nat. Methods* **10,** 999–1002 (2013).
13. Edgar, R. C. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10,** 996–998 (2013).
14. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13,** 581–583 (2016).
15. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci.* **102,** 11070–11075 (2005).
16. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13780–5 (2007).
17. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12,** (2011).
18. Faust, K. & Raes, J. Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10,** 538–550 (2012).
19. Faust, K. *et al.* Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput. Biol.* **8,** (2012).
20. Zhang, Z. *et al.* Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *ISME J.* **8,** 881–893 (2014).
21. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: Unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25,** 56–66 (2015).
22. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).
23. Gibbons, S. M., Kearney, S. M., Smillie, C. S. & Alm, E. J. Two dynamic regimes in the human gut microbiome. *PLoS Comput. Biol.* **13,** 1–20 (2017).
24. Tropini, C., Earle, K. A., Huang, K. C. & Sonnenburg, J. L. The Gut Microbiome : Connecting Spatial Organization to Function. *Cell Host Microbe* **21,** 433–442 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

25. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. comparison with gene expression arrays RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. 1509–1517 (2008). doi:10.1101/gr.079558.108
26. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
27. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science (80-.).* **352,** 560–564 (2016).
28. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163,** 1079–1095 (2015).
29. Almonacid, D. E. *et al.* 16S rRNA gene sequencing and healthy reference ranges for 28 clinically relevant microbial taxa from the human gut microbiome. *PLoS One* **12,** 1–15 (2017).
30. Stämmler, F. *et al.* Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4,** 1–13 (2016).
31. Elowitz, M. B., Levine, A. J. & Siggia, E. D. Stochastic Gene Expression in a Single Cell. *Science (80-.).* **297,** 1183–1187 (2002).
32. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS* **99,** (2002).
33. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science (80-.).* **304,** 1811–1814 (2004).
34. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).
35. Stein, R. R. *et al.* Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Comput. Biol.* **9,** 31–36 (2013).
36. Martin-Platero, A. M. *et al.* High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat. Commun.* **9,** 266 (2018).
37. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci.* **110,** 12804–12809 (2013).
38. Sonnenburg, E. D. *et al.* Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141,** 1241–1252 (2010).
39. Sonnenburg, E. D. *et al.* Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529,** 212–215 (2016).
40. Rakoff-Nahoum, S., Foster, K. R. & Comstock, L. E. The evolution of cooperation within the gut microbiota. *Nature* **533,** 255–259 (2016).
41. Rakoff-Nahoum, S., Coyne, M. J. & Comstock, L. E. An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24,** 40–49 (2014).
42. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79,** 5112–5120 (2013).
43. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. 1–3 (2010).
44. Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. & Al, W. E. T. Naı ¨ ve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. **73,** 5261–5267 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

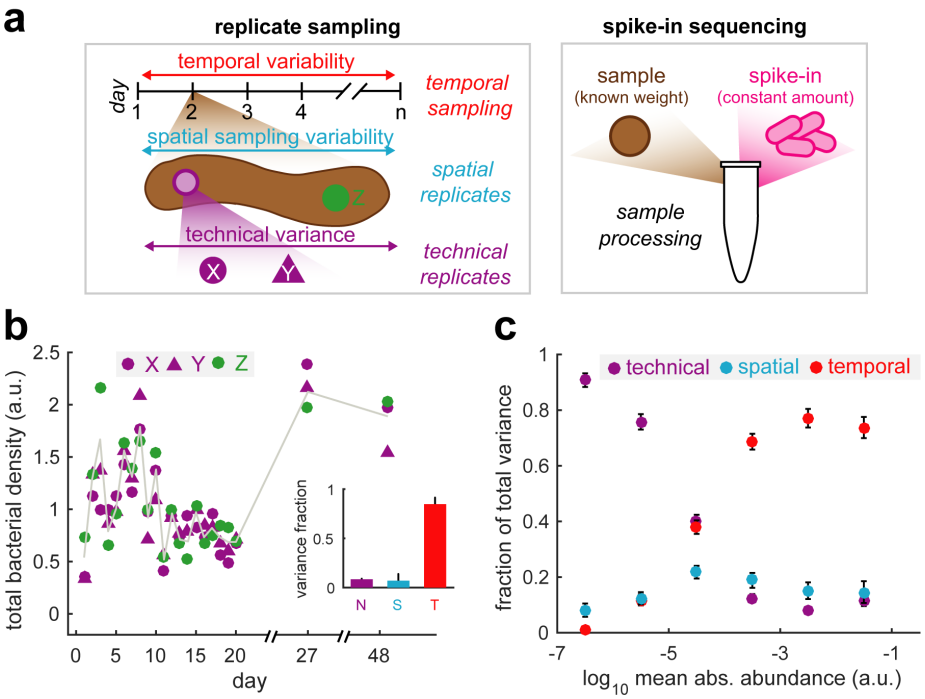

Figure 1: Variance decomposition of absolute gut bacterial abundances. a) Schematic of replicate sampling protocol and spike-in procedure. **b)** Total bacterial densities in the human gut microbiome over time and different sampling locations. X and Y correspond to technical replicate measurements of total bacterial density from a single spatial location while Z corresponds to a second spatial replicate. Gray line indicates the mean of spatial replicates. Total densities are reported in arbitrary units and normalized to a mean of unity (see Methods). Inset, variance in total bacterial loads attributed to technical (N, purple), spatial (S, blue) and temporal (T, red) factors as calculated by the variance decomposition model. **c)** Variance decomposition of individual OTU abundances. Absolute OTU abundances were obtained by multiplying relative abundance profiles by the total bacterial density in each sample and are reported in arbitrary units (see Methods). OTUs were binned by their mean abundance across all samples with data showing the average variance contribution of technical, spatial and temporal sources to OTUs within each bin. Error bars denote the standard error of the mean.

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

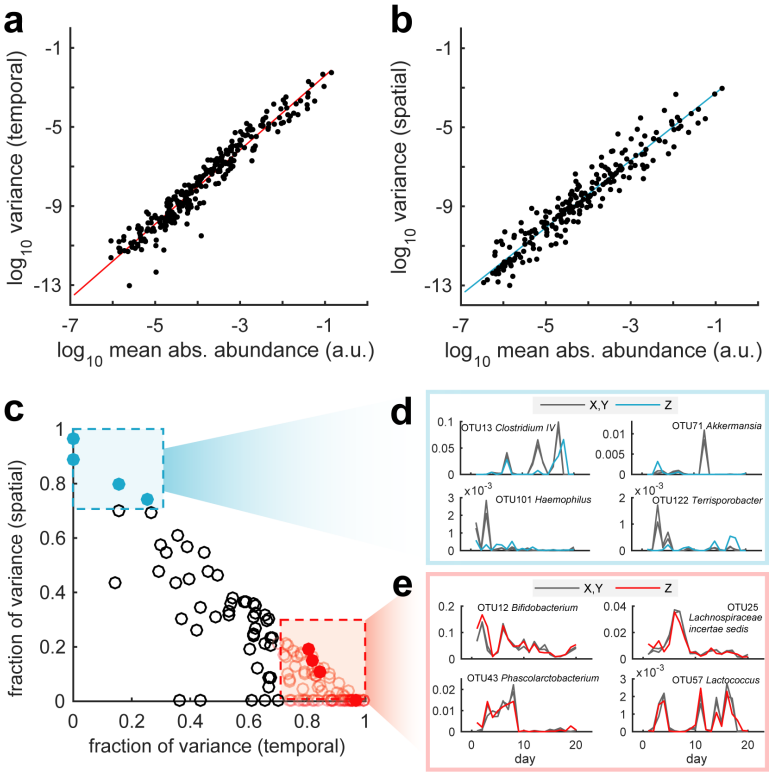

Figure 2: Temporal and spatial variation of human gut microbiota. a) Temporal and **b)** spatial variances of human gut microbiota abundances as predicted by the variance decomposition model. Data points correspond to the mean and model estimated temporal or spatial variance of individual OTUs. Power law coefficients were estimated with regression on log-transformed axes (regression slopes b=1.9, b=1.7 for temporal and spatial variability respectively) **c)** Identification of OTUs with high temporal or spatial variance. Boxes indicate strains with a contribution of temporal (red) or spatial (blue) variance relative to total variance exceeding 0.7. Only highly abundant OTUs ($\log_{10}$ mean absolute abundance > -4) are shown. **d,e)** Time series of individual OTUs whose abundance variation are comprised predominantly of temporal (red) or spatial sources (blue), indicated with filled circles in **c)**. Gray lines denote abundances of technical replicates (X,Y) obtained from the same spatial location and colored lines indicate abundances of the second spatial replicate (Z).

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

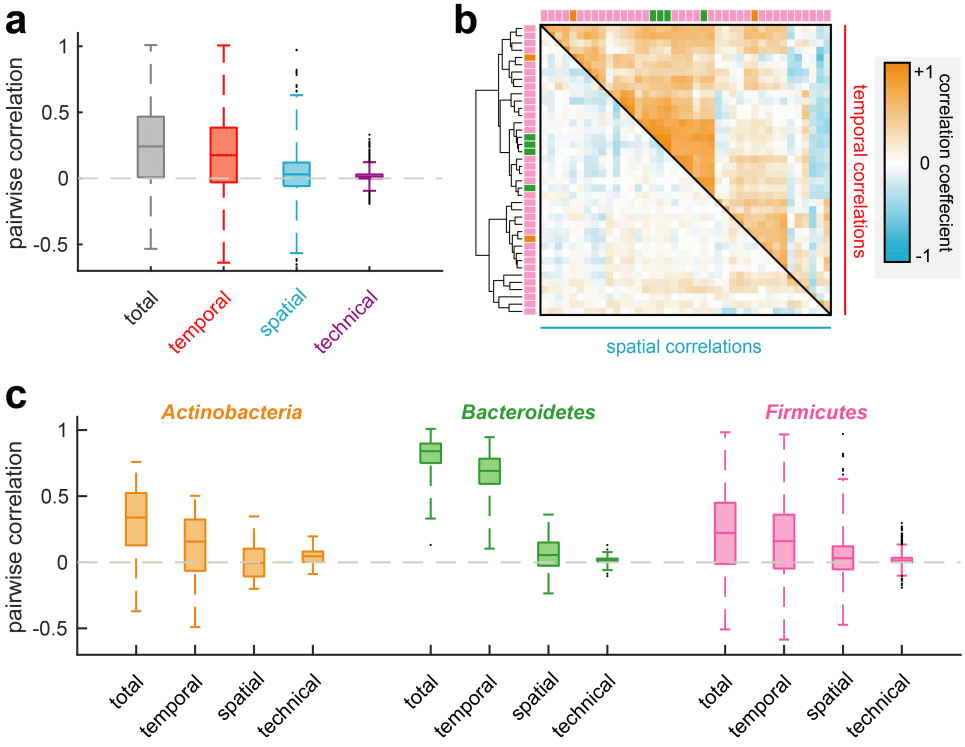

Figure 3: Decomposition of human gut microbiota abundance correlations. a) Boxplots of total, temporal, spatial and technical correlations for all pairs of highly abundant OTUs (log_{10} mean absolute abundance > -4). **b)** Pairwise temporal and spatial correlations for top 40 most abundant OTUs. OTUs are clustered by similarity in temporal correlation profiles and colored by phyla indicated in c). **c)** Pairwise OTU abundance correlations within three major phyla in the human gut microbiome (n=10,14, and 103 OTUs belonging to Actinobacteria, Bacteroidetes, and Firmicutes respectively). Correlations were restricted to pairs of highly abundant OTUs belonging to the indicated phyla.

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Material and methods

Ethical review

This study was approved and conducted under Columbia University Medical Center Institutional Review Board protocol AAAR0753. Written informed consent was obtained from the subject in the study, a healthy male adult.

Sample collection and storage

Fecal samples were collected by the subject daily over the course of twenty days, with two additional samples taken on days 27 and 48 of the study. After defecation, inverted sterile 200 µL pipette tips (Rainin RT-L200F) were used to core out a small sample from the stool, and placed immediately in a sterile cryovial (Sarstedt 72.694.106). On each day of stool collection, two samples were obtained from independent spatial locations at least >5cm in distance on the same stool specimen. Samples were immediately placed in a -20C freezer and transferred to a -80C freezer for long term storage.

Spike-in strain for calculation of bacterial absolute abundance

A spike-in approach was utilized during sample processing to allow for calculation of bacterial absolute abundance per mass of fecal matter. Sporocarcina pasteurii (ATCC 11859), an environmental bacterium not found in human feces, was grown to saturation in NH4-YE medium (ATCC medium 1376) and concentrated by centrifugation, resuspended in ~0.1X volume phosphate buffered saline with 20% glycerol and stored in cryovials at -80C for subsequent use during genomic DNA extraction.

Replicate sampling experimental protocol

To enable decomposition of gut bacterial abundance variability into temporal, spatial and technical contributions, we utilized a replicate sampling approach. Specifically, on each day, two fecal samples were collected from independent spatial locations on the same stool specimen. For one of these samples, two technical replicates were prepared in parallel by splitting the individual fecal core. Thus, a total of three samples were processed for each day of the time series: two technical replicates from a single spatial location (denoted samples X and Y) and a second spatial replicate (sample Z). To characterize spatial variaton at a single time point in further detail, we performed a more extensive set of sampling experiments in which fecal samples were collected from 14 independent locations >2 cm in distance on the same stool specimen. One of these samples was further subjected to 12 independent rounds of sample processing and sequencing to directly assess technical variability. Metadata associated with all samples can be found in Table S1. Theoretical details associated with the DIVERS approach can be found in the Supplementary Note.

Sample genomic DNA extraction

Genomic DNA (gDNA) extraction was performed using a custom liquid handling protocol based on the Qiagen MagAttract PowerMicrobiome DNA/RNA Kit (Qiagen 27500-4-EP) adapted for lower volumes. Briefly, a 96 well plate (Axygen P-DW-20-C) was loaded with 1mL of 0.1mm Zirconia Silica beads (Biospec 11079101Z) using a loading device (Biospec 702L). During sample processing, appropriate negative controls were run on each plate (i.e. water control). 10uL of thawed and concentrated spike-in strain was added to each well. 10-100mg of each sample

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(average 45.9 mg, standard deviation 14.7 mg) was added to the plate using a sterile plastic spatula, and the weight added for each sample was determined via an analytical balance. 750 µL of lysis solution was then added to each well (90 mL master mix, 9 mL 1M Tris HCl pH 7.5, 9 mL 0.5M EDTA pH 8.0, 11.25 mL 10% SDS, 22.5 mL Qiagen lysis reagent, 38.25 mL nuclease free water). The plate was then centrifuged down for 1 m at 4500 g and a bead sealing mat was affixed to the plate (Axygen AM-2ML-RD). The plate was then placed on a bead beater (Biospec 1001) and subjected to bead beating for 5 m followed by 10 m for cooling. This bead beating cycle was repeated, for a total of 10 m of bead beating. The plate was then centrifuged down for 5 m at 4500 g and 200uL of supernatant was transferred to a V-bottom microplate. 35 µL of Qiagen inhibitor removal solution was added to each well and mixed by vortexing, incubated 4C for 5 m, and the plate was again centrifuged down for 5 m at 4500 g. 100 µL of supernatant was removed from the plate and placed in a round-bottom plate (Corning 3795). The plate was then placed on a robotic liquid handler (Biomek 4000) for magnetic bead purification of the supernatant per the manufacturers recommendations but at a scaled volume; magnetic beads in binding solution were mixed in each well, and subjected to 3 washes with wash solution and elution in 100uL of nuclease free water into a new plate.

16S rRNA amplicon sequencing

16S sequencing of the V4 region was performed utilizing a custom protocol and a dual indexing scheme adapted from Kozich et al[42]. Briefly, dual indexing sequencing primers were adapted from the previous study but we utilized Illumina Nextera barcode sequences and altered 16S primers to match updated 505f and 806rB primers (see Table S2 for sequences). A 20 µL PCR amplification was set up in a 96 well skirted PCR microplate (1 µM forward 5XX barcoded primer, 1 µM reverse 7XX barcoded primer, 1 µL prepared gDNA, 10uL NEBNext Q5 Hot Start HiFi Master Mix [NEB M0543L], 0.2X final concentration SYBR Green I). A quantiative PCR amplification (98C 30s; cycle: 98C 20s, 55C 20s, 65C 60s, 65C 5m) was performed and cycling was stopped during exponential amplification (typically 12-20 cycles) and the reaction was advanced to the final extension step.

The resulting PCRs were quantified utilizing a SYBR Green I dsDNA assay; 2 µL of PCR product was added to 198 µL of TE with 1X final concentration SYBR Green I and fluorescence was quantified on a microplate reader. Samples were pooled based on this quantification on a robotic liquid handler (Biomek 4000). The resulting ~390bp amplicon from the pool was then gel-purified utilizing a 2% E-gel (Invitrogen) and Wizard SV gel extraction kit.

Final libraries were then quantified by Qubit dsDNA HS assay and sequenced on the Illumina MiSeq platform (V2 500 cycle kit) according to the manufacturers instructions with modifications. Specifically, the library was loaded at 10 pM with 20% PhiX spike in, and custom sequencing primers were spiked into the MiSeq reagent cartridge (6uL of 100 µM stock; well 12: read1, well 13: index1, well 14: read2).

Sequence analysis and OTU clustering

Resulting sequence data was analyzed with the USEARCH[43] pipeline. Specifically, raw reads were merged using the –fastq_mergepairs command with options –fastq_maxdiffs 10 –fastq_maxdiffpct 10. Merged sequences were filtered using the –fastq_filter command with options –fastq_maxee 1.0 and –fastq_minlen 240. Resulting sequences were dereplicated (–derep_fulllength), clustered into OTUs (–cluster_otus) and the merged reads were searched against OTUs sequences (–usearch_global) at 97% identity. Taxonomic assignments were made to OTUs using the RDP classifier[44].

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Calculation of OTU absolute abundances

Total bacterial loads in each sample were calculated using the following formula:

$$R_i = \frac{C_0}{C_0 + \rho_i W_i}$$

Here, R_i is the sequenced relative abundance of the spike-in strain in sample i, C_0 is the constant amount of spike-in strain (units of total DNA copies) added to each sample, W_i is the weight of the fecal sample i (mg), and ρ_i is the total bacterial density per fecal mass (DNA copies/mg). Solving for ρ_i,

$$\rho_i = \frac{C_0(1 - R_i)}{R_i W_i}$$

where we have measured R_i and W_i experimentally. Note that relative changes in ρ_i are independent of the constant C_0, and therefore its exact number need not be known. We therefore scaled total bacterial densities across samples to a mean of unity. Relative abundance profiles (with the spike-in strain excluded) were then multiplied by this scaled quantity to obtain absolute OTU abundances in arbitrary units that were used for all analyses.

Variance and covariance decomposition of OTU abundances and total bacterial loads

Variances and covariances of OTU abundances were calculated across the twenty consecutive days of sampling. The variance decomposition of total bacterial densities also included samples taken from days 27 and 48 of the times series. To minimize artifacts of technical noise, only OTUs with a $\log_{10}$ mean absolute abundance > -4 were included when stratifying variance contribution profiles across phyla and in the covariance decomposition analysis. This cutoff was chosen based on the variance profiles of individual OTUs shown in Fig. 1c. 16S rRNA sequence distances were calculated using the pairwise2 module of Biopython for the comparison of correlation contributions across phyla.

Identification of OTUs with high temporal or spatial variance

To minimize effects of technical noise, OTUs were first filtered by abundance ($\log_{10}$ mean absolute abundance > -4). Of the remaining OTUs, those with temporal or spatial variance contributions exceeding 70% were identified and are listed in Table S3.

Removal of temporal or spatial variability from fecal samples

To ensure our model did not report any temporal or spatial variance contributions when those signals were absent from data, we conducted two sets of experiments specifically removing either temporal or spatial variability of OTU abundances from fecal samples. To eliminate temporal contributions, we re-sampled a single stool specimen ten times total to simulate five consecutive days of time series sampling. To eliminate spatial variability, replicate sampling was conducted for eight consecutive days. On each day fecal samples obtained from independent spatial locations were homogenized together by combining fecal samples and mechanically

bioRxiv preprint doi: https://doi.org/10.1101/310649; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

homogenizing in 1X phosphate buffered saline with a P200 pipette tip. The resulting homogenized sample was then split into technical triplicates and procesed as per the normal DIVERS protocol.