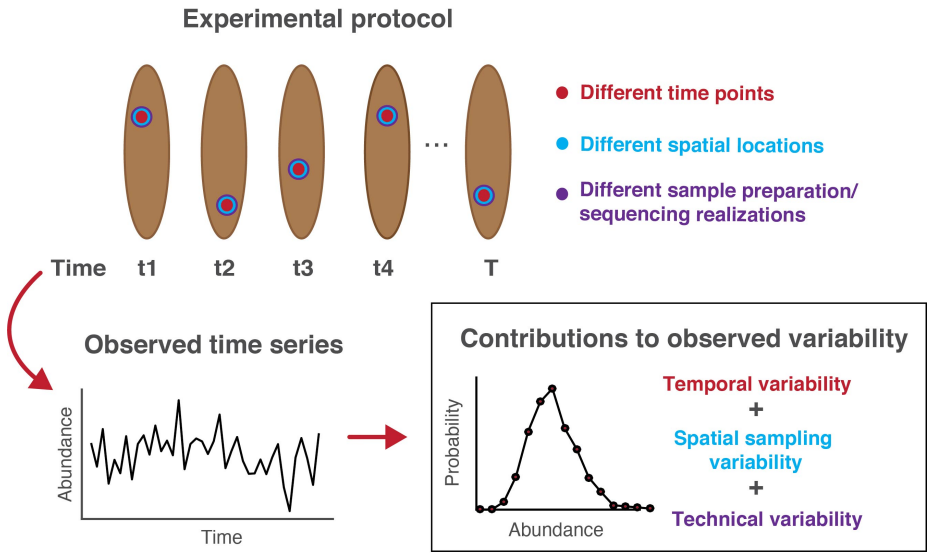
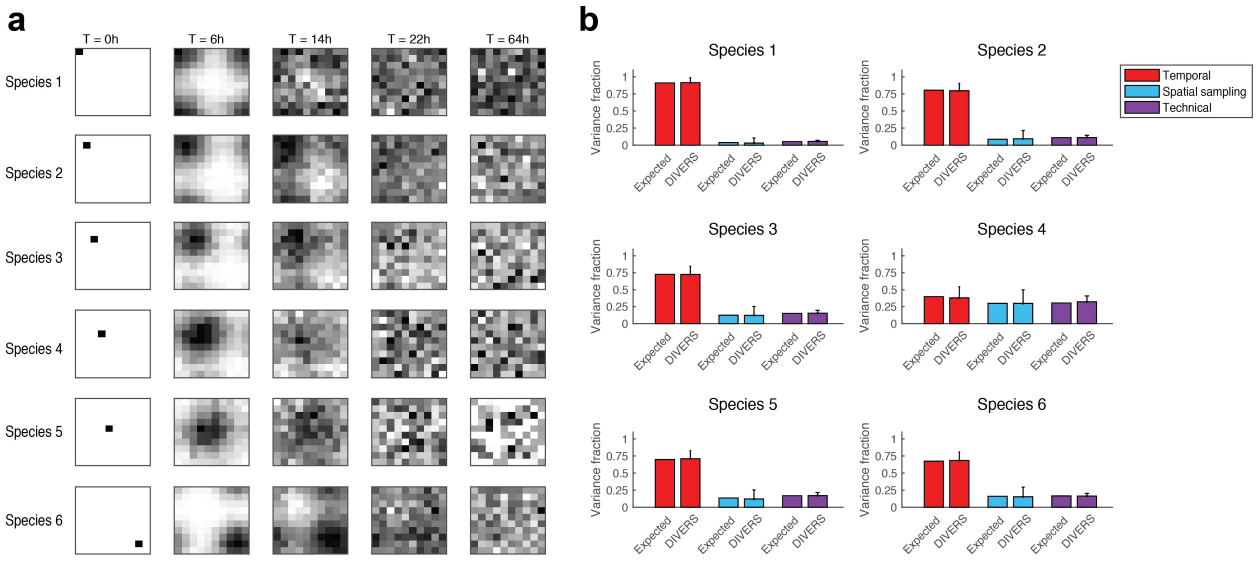
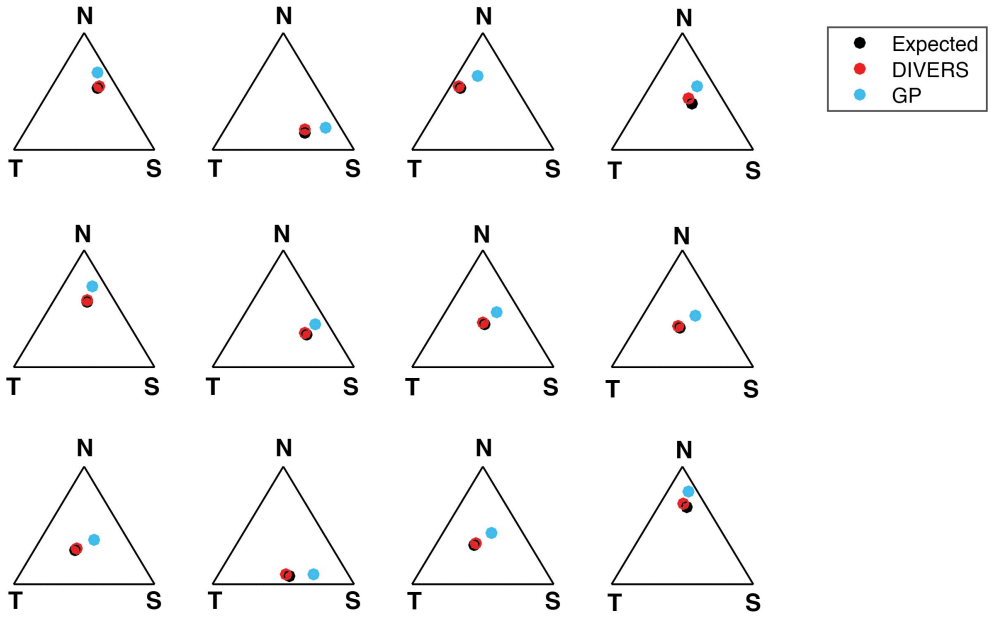
Supplementary Figures



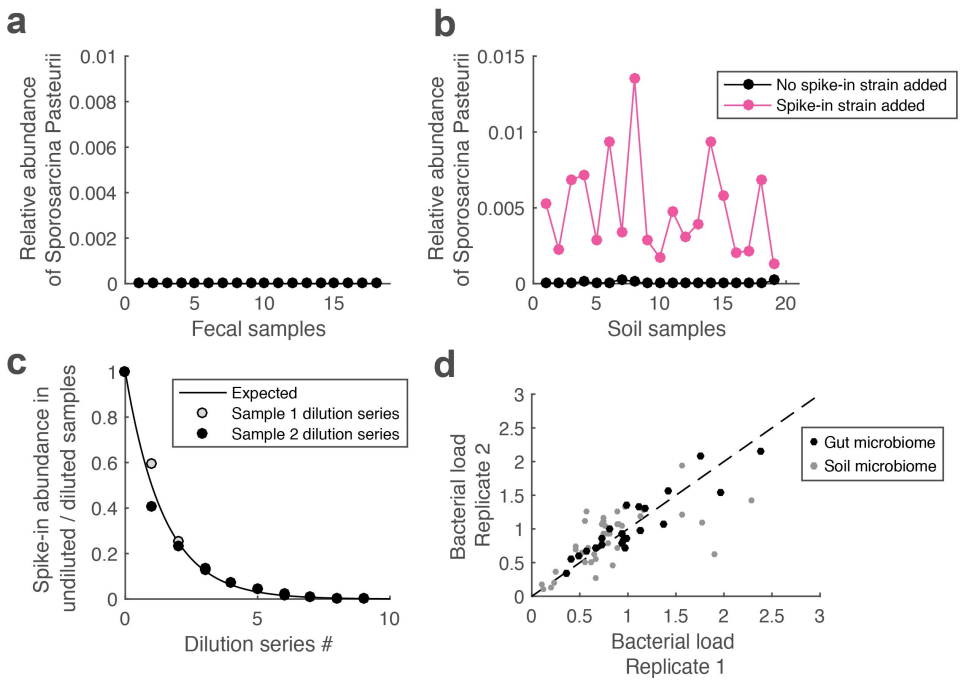
Supplementary Figure 1. Representative sampling protocol for longitudinal studies of the gut microbiome. Observed time series of individual taxa are expected to reflect true temporal dynamics. However, time series traces may also reflect fecal spatial heterogeneity associated with sampling different spatial locations on stool at different time points. Observed variability may also reflect technical noise associated with sample preparation and sequencing at each time point. Therefore, using this sampling protocol, it is not possible to quantify the contributions of these three sources of variability to the total observed variability of individual bacterial abundances, although this knowledge is crucial for data analysis and interpretation. A similar problem exists for other microbiome studies, even when spatial sampling location can be controlled for, due to the effects of technical variability.



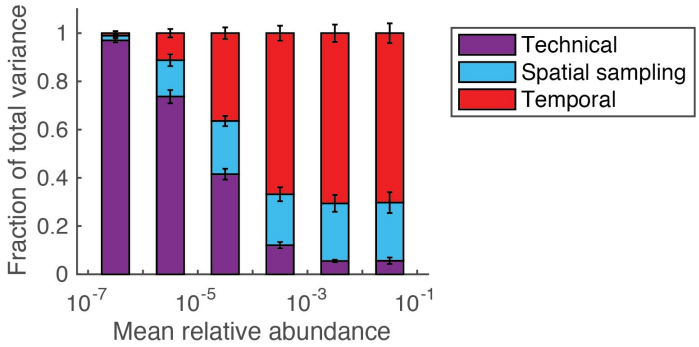
Supplementary Figure 2. Assessment of the DIVERS variance decomposition model using simulated bacterial dynamics. (a) Stochastic model of spatiotemporal dynamics in an interacting bacterial community. Simulations were performed on a 10 x 10 lattice with continuous boundary conditions, where dynamics of each species were governed by birth events, death events or random migration to neighboring locations. Species interactions were modeled using density-dependent logistic growth at each location (Online Methods). Results for six species with non-zero steady state abundances are shown. **(b)** Expected contributions of temporal, spatial sampling and technical sources to total abundance variance, and predictions from the DIVERS variance decomposition model. Technical variability was modeled using Poisson sampling noise centered on the true abundances at each spatial location (Online Methods). Results for six species in the simulated community are shown and correspond to those in panel a. Expected variance contributions were empirically calculated from simulated data (Online Methods). Both expected and DIVERS variance contributions were calculated across thirty time points after neglecting any initial transient behavior. Error bars represent standard deviations based on n = 5,000 re-samplings from different pairs of spatial locations in the environment.



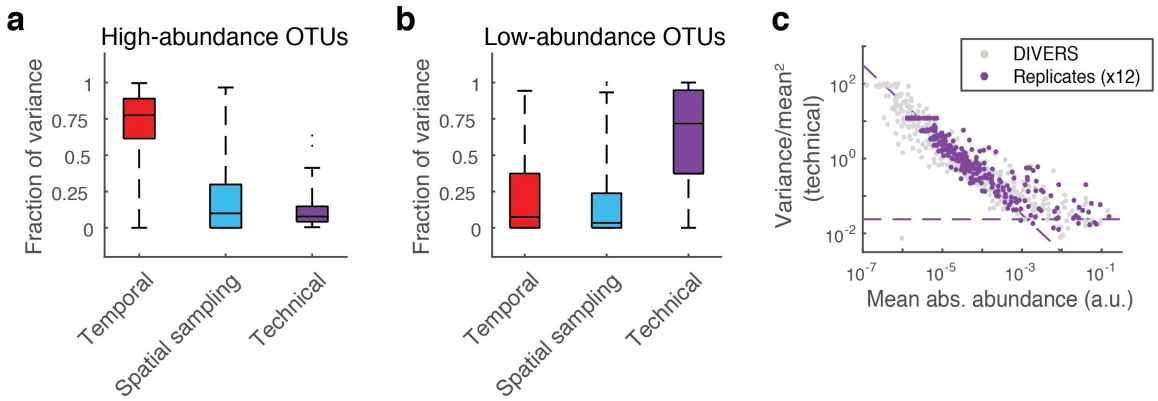
Supplementary Figure 3. Comparison of DIVERS to the Gaussian process variance decomposition model. The speed and accuracy of DIVERS was compared a recent approach where Gaussian process decomposition was applied to study microbiome variability. True temporal (T), spatial (S) and technical (N) variance contributions were used as inputs into a generative statistical model to simulate microbiota abundance dynamics (Online Methods). Results from twelve representative simulations are shown as ternary plots. The true set of variance contributions are represented by black dots, while DIVERS estimates are represented by red dots, and estimates from the Gaussian process decomposition model are represented by blue dots. DIVERS was significantly more accurate than the Gaussian process decomposition model (r.m.s. error = 0.017 for DIVERS, r.m.s. error = 0.11 for Gaussian process decomposition; $p = 2.5 \times 10^{-34}$, Wilcoxon test based on 100 random sets of true variance fraction contributions). Notably, DIVERS variance decomposition required less than a second of computing time on a standard laptop computer, while the Gaussian process inference procedure required roughly four minutes for each individual OTU.



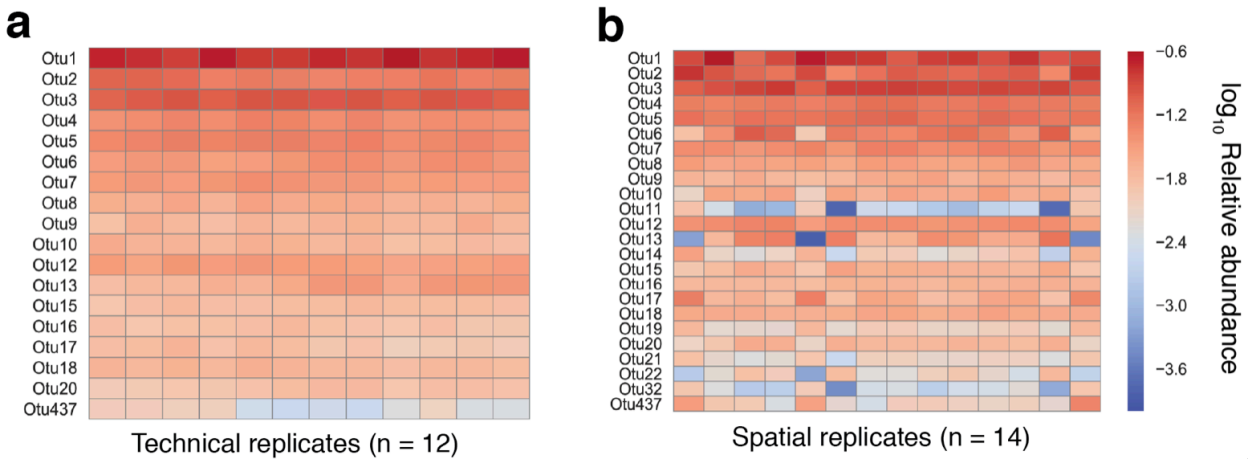
Supplementary Figure 4. Validation of the spike-in sequencing approach to estimate total bacterial loads in collected fecal and soil samples. (**a**) Relative endogenous abundances of the spike-in strain, *Sporosarcina Pasteurii*, in eighteen fecal samples collected from the same human individual presented in the main text. No reads belonging to the spike-in strain were detected in any samples. (**b**) Relative endogenous abundances of *Sporosarcina Pasteurii* across nineteen of the soil sampling sites in Central Park. Observed relative abundances are consistent with sample read-through (contamination from other samples in the same sequencing run) and are on average two orders of magnitude lower than sequenced abundances in the same samples for which the spike-in strain was added. (**c**) Expected and observed behavior of spike-in strain abundances across two serially diluted fecal samples (Online Methods). Expected behavior was derived based on the dilution factor of 2 used in the dilution series of each original fecal sample (Online Methods). (**d**) Technical replicate measurements of bacterial loads in fecal and soil samples. Measurements are separately normalized to a mean of one within fecal or soil samples. Technical replicates across all fecal and soil samples are highly correlated (Pearson's r = 0.82, n = 76 samples with two technical replicates each).



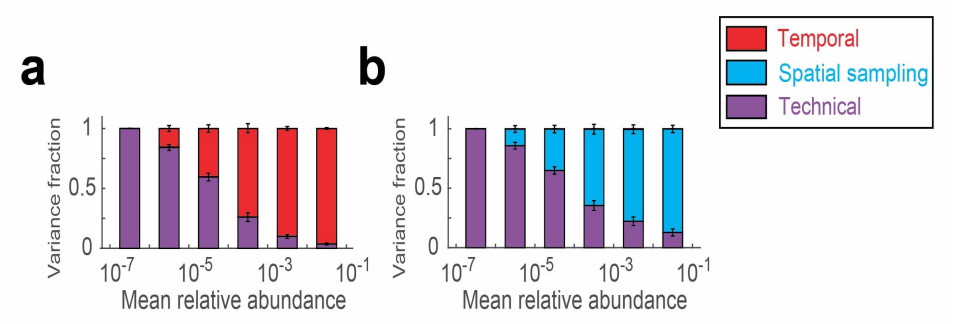
Supplementary Figure 5. Variance decomposition of OTU relative abundances in the human gut microbiome. OTUs are binned by mean relative abundance across samples. Stacked bars indicate the average fraction of total variance attributed to temporal, spatial sampling and technical sources for OTUs within each bin. A total of n = 433 OTUs were used. Error bars denote the SEM.



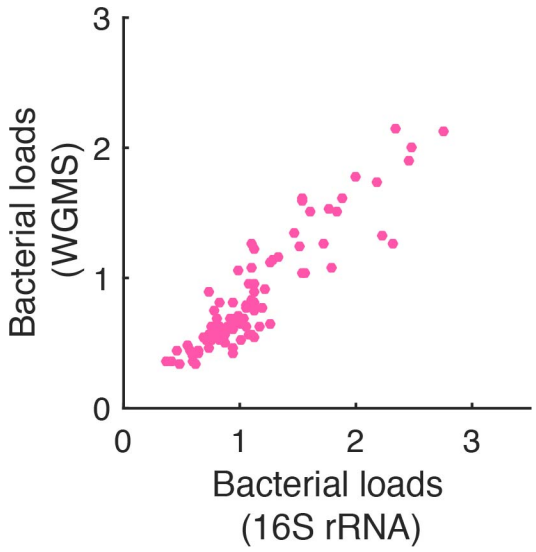
Supplementary Figure 6. Variance decomposition for high and low-abundance OTUs in the human gut microbiome. Temporal, spatial sampling and technical contributions to total variance for all OTUs with (**a**) mean absolute abundance > 10^{-4} (n = 133 OTUs) and (**b**) mean absolute abundance < 10^{-4} (n = 300 OTUs). Boxes show the median and interquartile ranges, with maximum whisker lengths three times the interquartile range. **(c)** OTU abundance variability due to technical noise. n=12 independent technical replicates were processed from fecal samples obtained from a single spatial location of a stool specimen. Purple dots show the normalized technical variability (variance/mean^2) as a function of average abundance across twelve technical replicates. Technical noise profiles obtained from DIVERS are shown in gray dots. The inverse scaling expected from Poissonian sampling noise is indicated with the dashed line with slope = -1. A noise floor is observed at high OTU abundances (indicated by the horizontal dashed line) as a result of variability in the spike-in process to estimate total bacterial loads.



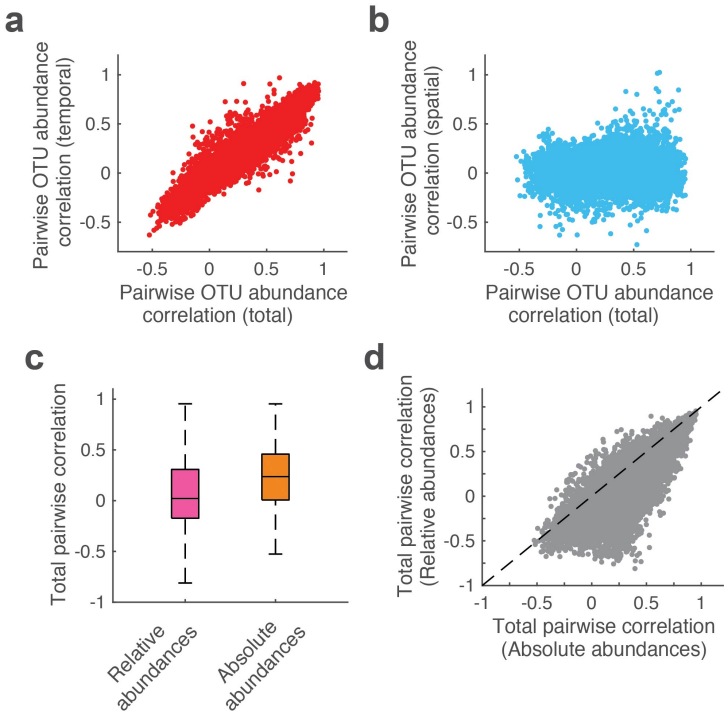
Supplementary Figure 7. Technical noise and spatial sampling variability of OTUs in the human fecal microbiome. (a,b) Visualization of individual OTU relative abundance across multiple technical and spatial replicates. Technical replicates (n=12) were obtained by subjecting the same underlying fecal matter to multiple rounds of sample preparation and 16S rRNA sequencing. Spatial replicates (n=14) were collected from multiple random locations of a single defecated stool specimen. Colors denote the $\log_{10}$ relative abundance of OTUs in each sample. Only the top most abundant OTUs across each set of replicates are shown. Spatial replicate variability suggests that fecal heterogeneity is an inherent feature of fecal bacterial sequencing studies.



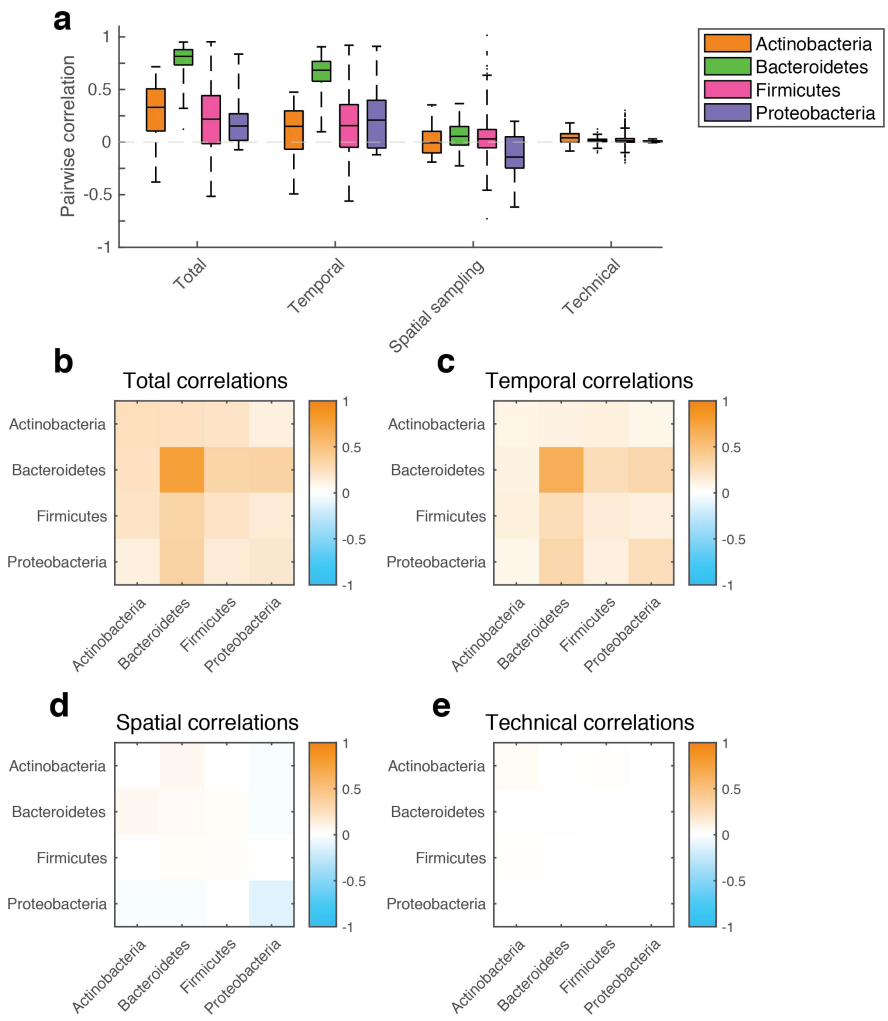
Supplementary Figure 8. Variance decomposition of microbiota abundances from control experiments. (a) DIVERS applied to stool samples without spatial variability; **(b)** DIVERS applied to stool samples without temporal variability. Stacked bars indicate the average fraction of total variance attributed to temporal, spatial sampling and technical sources for OTUs within each bin. A total of n = 314 OTUs were used. Error bars denote the SEM.



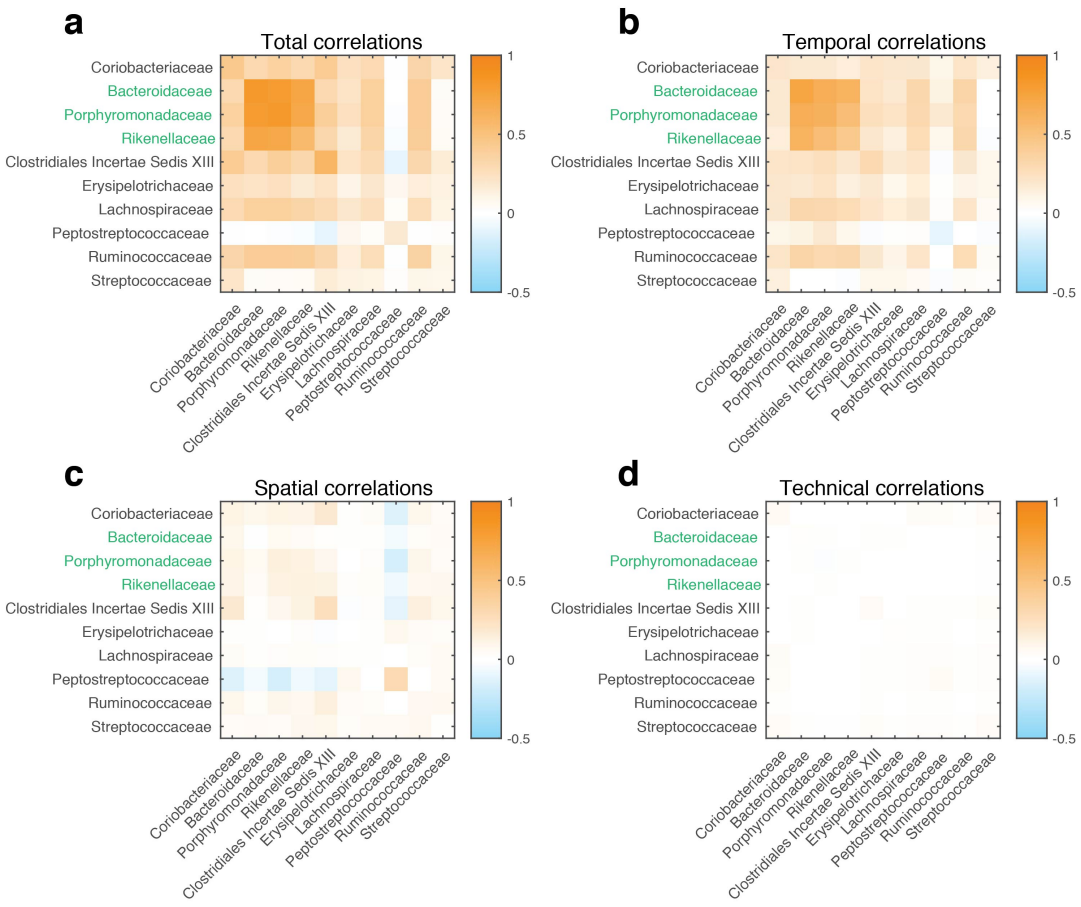
Supplementary Figure 9. Correlation between fecal bacterial loads between 16S rRNA or whole-metagenome shotgun sequencing. Correlation between fecal bacterial loads estimated using the spike-in approach in conjunction with either 16S rRNA amplicon sequencing (X-axis) or whole-metagenome shotgun sequencing (Y-axis); n = 88, Pearson's r = 0.91.



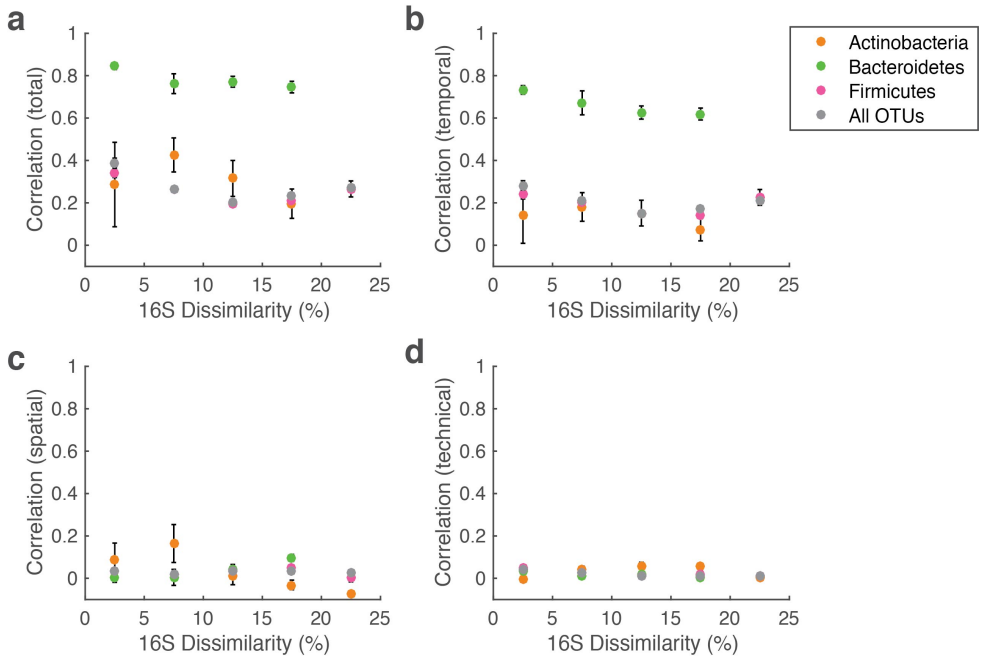
Supplementary Figure 10. Pairwise OTU abundance correlations in the human gut microbiome. Relationship between total pairwise OTU absolute abundance correlations and (**a**) temporal and (**b**) spatial abundance correlations across all pairs of abundant OTUs (mean absolute abundance > 10^{-4}). There is a significant correlation between temporal and total correlations (Pearson's r = 0.94, p < 1e-10). (**c**) Total correlations calculated across all pairs of highly abundant OTUs (mean absolute abundance > 10^{-4}) using relative (pink) or absolute (orange) abundances. Boxes show the median and interquartile ranges, with maximum whisker lengths three times the interquartile range. (**d**) Total pairwise OTU abundance correlations calculated using absolute abundances versus pairwise correlations using relative abundances; each point represents a pair of OTUs. Dashed line indicates the y = x line. n = 133 OTUs were used in the calculations.



Supplementary Figure 11. Correlations of OTU abundances within and between different phyla in the human gut. (**a**) Boxplots of total, temporal, spatial and technical correlations of OTU abundances, where pairwise comparisons were made between OTUs within the indicated phyla. Boxes show median and interquartile ranges, with maximum whisker lengths three times the interquartile range (n = 133 OTUs were used). (**b-e**) Average correlations of OTU abundances within and between different phyla. Colors indicate the average (**b**) total, (**c**) temporal, (**d**) spatial and (**e**) technical correlations between pairs of OTUs belonging to the indicated phyla. See online methods Eq. 9 for the definition of correlations.

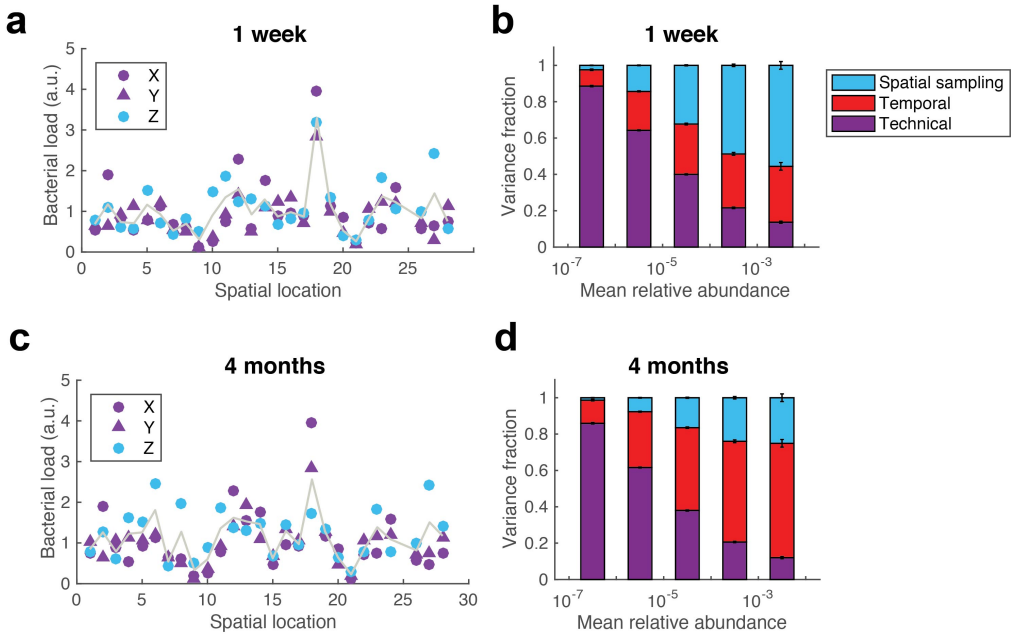


Supplementary Figure 12. Average correlations of OTU abundances within and between different microbial families in the human gut. Colors indicate the average (**a**) total, (**b**) temporal, (**c**) spatial and (**d**) technical correlations between pairs of OTUs belonging to the indicated families. The three families belonging to the Bacteroidetes phylum are shown in green. n = 133 OTUs were used. See online methods Eq. 9 for the definition of correlations.

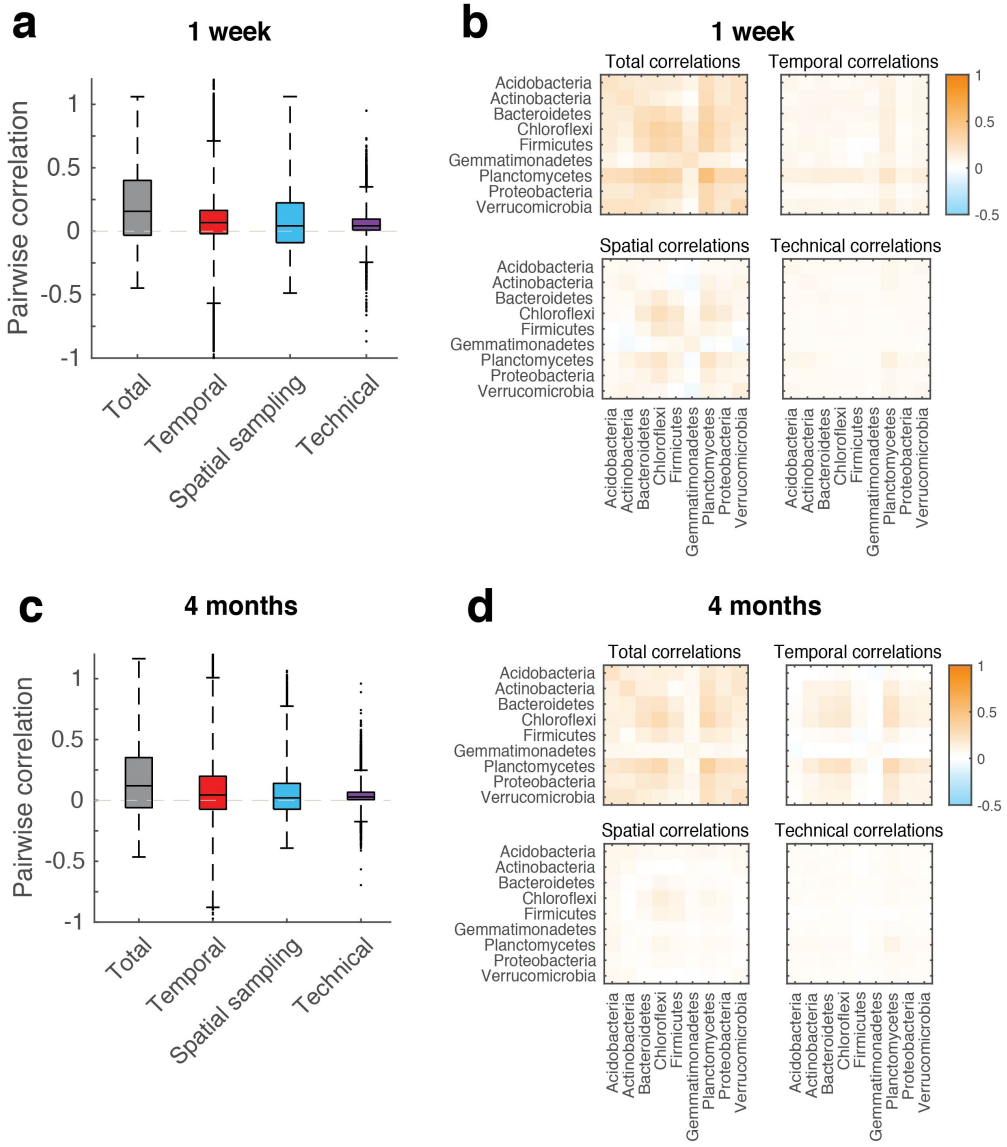


Supplementary Figure 13. Pairwise abundance correlations of gut bacterial OTU abundances within phyla at different phylogenetic distances. (**a**) Total, (**b**) temporal, (**c**) spatial and (**d**) technical correlations were calculated for pairs of OTUs belonging to the same phyla, but with different degrees of 16S rRNA sequence dissimilarity (n = 133 OTUs were used). OTU pairs are binned by 16S sequence dissimilarity and mean correlations within each bin are shown for the indicated phyla with error bars denoting the SEM. Pairwise correlations for all abundant OTUs are shown in gray. Proteobacteria were excluded from the analysis due to insufficient sample size.

Supplementary Figure 14. Sampling sites from which soil bacteria were collected for 16S rRNA sequencing. Sites were located around the periphery of a small pond in Central Park, Manhattan.



Supplementary Figure 15. Application of DIVERS to a Central Park soil community. (**a,c**) Bacterial loads of soil samples across different spatial locations. X and Y correspond to measurements made at a single time point, while Z corresponds to measurements from the second time point collected either one week or four months apart. (**b,d**) Variance decomposition of n = 24667 individual OTU relative abundances. OTUs are binned by their mean relative abundance across all samples, and stacked bars show the average variance contribution of technical, spatial sampling and temporal sources to OTUs within each bin. Error bars represent the SEM. Temporal variability reflects average changes in the community at the indicated time scale. The gray line is the average ½(Z + ½(X+Y)) of the three bacterial loads.



Supplementary Figure 16. Decomposition of pairwise OTU abundance correlations in the Central Park soil microbiome. (**a,c**) Boxplots of total, temporal, spatial and technical correlations for all pairs of abundant OTUs (average $\log_{10}$ absolute abundance > -3.5) in data collected one week and four months apart. Boxes denote the median and interquartile ranges, with maximum whisker lengths three times the interquartile range. (**b,d**) Decomposition of pairwise OTU abundance correlations within and between different phyla in data collected one week and four months apart. Heatmaps show the average total, temporal, spatial and technical correlations between pairs of OTUs belonging to the indicated phyla. n = 68, 66, 115, 22, 8, 8, 18, 204, and 36 OTUs of Acidobacteria, Actinobacteria, Bacteroidetes, Chloroflexi, Firmicutes, Gemmatimonadetes, Planctomycetes, Proteobacteria, and Verrucomicrobia respectively in panel **(a,b)** and **(c,d)**. See online methods for the definition of correlations.

Supplementary Table 1. Metadata for all fecal samples.

Sample Name	Day	Time	Spatial_Rep.	Tech._Rep.	Weight (mg)
d16s1r1	1	9:00AM	1	1	41.5
d16s2r1	1	9:00AM	2	1	85.5
d16s2r2	1	9:00AM	2	2	61.6
d17s1r1	2	7:00PM	1	1	49.5
d17s2r1	2	7:00PM	2	1	56.3
d17s2r2	2	7:00PM	2	2	52.2
d18s1r1	3	3:00PM	1	1	46.5
d18s2r1	3	3:00PM	2	1	63
d18s2r2	3	3:00PM	2	2	73.9
d19s1r1	4	7:00PM	1	1	57.4
d19s2r1	4	7:00PM	2	1	37.1
d19s2r2	4	7:00PM	2	2	37.5
d20s1r1	5	12:00PM	1	1	46.4
d20s2r1	5	12:00PM	2	1	39.6
d20s2r2	5	12:00PM	2	2	29.3
d21s1r1	6	10:30PM	1	1	49.9
d21s2r1	6	10:30PM	2	1	52.4
d21s2r2	6	10:30PM	2	2	41.5
d22s1r1	7	11:56PM	1	1	47.7
d22s2r1	7	11:56PM	2	1	36.7
d22s2r2	7	11:56PM	2	2	35.8
d23s1r1	8	4:00PM	1	1	43
d23s2r1	8	4:00PM	2	1	41
d23s2r2	8	4:00PM	2	2	44.4
d24s1r1	9	4:30PM	1	1	64.9
d24s2r1	9	4:30PM	2	1	38.8
d24s2r2	9	4:30PM	2	2	35.3
d25s1r1	10	11:00PM	1	1	59.2
d25s2r1	10	11:00PM	2	1	39.2
d25s2r2	10	11:00PM	2	2	56.4
d26s1r1	11	10:00AM	1	1	36.7
d26s2r1	11	10:00AM	2	1	58.8
d26s2r2	11	10:00AM	2	2	68.2
d28s1r1	13	3:30PM	1	1	57.4
d28s2r1	13	3:30PM	2	1	56.2
d28s2r2	13	3:30PM	2	2	45.5
d29s1r1	14	3:00PM	1	1	58
d29s2r1	14	3:00PM	2	1	53
d29s2r2	14	3:00PM	2	2	64.8
d30s1r1	15	4:00PM	1	1	71.4
d30s2r1	15	4:00PM	2	1	16.7
d30s2r2	15	4:00PM	2	2	23
d31s1r1	16	8:30PM	1	1	32.4
d31s2r1	16	8:30PM	2	1	63.5
d31s2r2	16	8:30PM	2	2	50.3
d32s1r1	17	5:00PM	1	1	62.3
d32s2r1	17	5:00PM	2	1	46.7
d32s2r2	17	5:00PM	2	2	68
d33s1r1	18	12:45PM	1	1	42.6
d33s2r1	18	12:45PM	2	1	38.1
d33s2r2	18	12:45PM	2	2	31.8
d34as1r1	19a	9:00AM	1	1	41.7
d34as2r1	19a	9:00AM	2	1	46.4
d34as2r2	19a	9:00AM	2	2	35.2

d34bs1r1	19b	7:00PM	1	1	57.1
d34bs2r1	19b	7:00PM	2	1	46.3
d34bs2r2	19b	7:00PM	2	2	46.2
d35s1r1	20	4:00PM	1	1	61.9
d35s2r1	20	4:00PM	2	1	44.6
d35s2r2	20	4:00PM	2	2	49
d42s1r1	27	7:30PM	1	1	58.3
d42s2r1	27	7:30PM	2	1	47.7
d42s2r2	27	7:30PM	2	2	54.9
d63s1r1	48	10:00PM	1	1	49.5
d63s2r1	48	10:00PM	2	1	38.2
d63s2r2	48	10:00PM	2	2	28.4
d30s3r1	15	4:00PM	3	1	60.8
d30s4r1	15	4:00PM	4	1	39.5
d30s5r1	15	4:00PM	5	1	44.7
d30s6r1	15	4:00PM	6	1	32.1
d30s7r1	15	4:00PM	7	1	54.1
d30s8r1	15	4:00PM	8	1	85.1
d30s9r1	15	4:00PM	9	1	39.7
d30s10r1	15	4:00PM	10	1	46.7
d30s11r1	15	4:00PM	11	1	65
d30s12r1	15	4:00PM	12	1	34.7
d30s13r1	15	4:00PM	13	1	49
d30s14r1	15	4:00PM	14	1	72.7
d30s2r3	15	4:00PM	2	3	26.4
d30s2r4	15	4:00PM	2	4	21.1
d30s2r5	15	4:00PM	2	5	29.1
d30s2r6	15	4:00PM	2	6	23.6
d30s2r7	15	4:00PM	2	7	15.6
d30s2r8	15	4:00PM	2	8	20.9
d30s2r9	15	4:00PM	2	9	11.8
d30s2r10	15	4:00PM	2	10	21
d30s2r11	15	4:00PM	2	11	16.4
d30s2r12	15	4:00PM	2	12	11.2

Supplementary Table 2. 16S sequencing primers utilized in the study.

Primer Name	Sequence (5' to 3')
16Sf_501	AATGATACGGCGACCACCGAGATCTACAC TAGATCGC TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_502	AATGATACGGCGACCACCGAGATCTACAC CTCTCTAT TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_503	AATGATACGGCGACCACCGAGATCTACAC TATCCTCT TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_504	AATGATACGGCGACCACCGAGATCTACAC AGAGTAGA TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_505	AATGATACGGCGACCACCGAGATCTACAC GTAAGGAG TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_506	AATGATACGGCGACCACCGAGATCTACAC ACTGCATA TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_507	AATGATACGGCGACCACCGAGATCTACAC AAGGAGTA TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sf_508	AATGATACGGCGACCACCGAGATCTACAC CTAAGCCT TATGGTAATT GT GTGYCAGCMGCCGCGGTAA
16Sr_701	CAAGCAGAAGACGGCATACGAGAT TCGCCTTA AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_702	CAAGCAGAAGACGGCATACGAGAT CTAGTACG AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_703	CAAGCAGAAGACGGCATACGAGAT TTCTGCCT AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_704	CAAGCAGAAGACGGCATACGAGAT GCTCAGGA AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_705	CAAGCAGAAGACGGCATACGAGAT AGGAGTCC AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_706	CAAGCAGAAGACGGCATACGAGAT CATGCCTA AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_707	CAAGCAGAAGACGGCATACGAGAT GTAGAGAG AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_708	CAAGCAGAAGACGGCATACGAGAT CCTCTCTG AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_709	CAAGCAGAAGACGGCATACGAGAT AGCGTAGC AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_710	CAAGCAGAAGACGGCATACGAGAT CAGCCTCG AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_711	CAAGCAGAAGACGGCATACGAGAT TGCCTCTT AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16Sr_712	CAAGCAGAAGACGGCATACGAGAT TCCTCTAC AGTCAGTCAG CC GGACTACNVGGGTWTCTAAT
16S_read1	TATGGTAATTGTGTGYCAGCMGCCGCGGTAA
16S_read2	AGTCAGTCAGCCGGACTACNVGGGTWTCTAAT
16S_index1	ATTAGAWACCCBNGTAGTCCGGCTGACTGACT

Supplementary Table 3. Metadata for all soil samples.

Sample Name	Soil Sample ID	Location	Time point	Site No.	Tech. Rep.	Biol. Rep.	Date	Time	Sample Weight (mg)
Soil1	T1S1	The Pool in Central Park, Manhattan	1	1	0	0	15-Jun-18	9:30AM-11:45AM	261.5
Soil2	T1S2	The Pool in Central Park, Manhattan	1	2	1	0	15-Jun-18	9:30AM-11:45AM	258.1
Soil3	T1S2	The Pool in Central Park, Manhattan	1	2	2	0	15-Jun-18	9:30AM-11:45AM	234.1
Soil4	T1S3	The Pool in Central Park, Manhattan	1	3	0	0	15-Jun-18	9:30AM-11:45AM	266.1
Soil5	T1S4	The Pool in Central Park, Manhattan	1	4	1	0	15-Jun-18	9:30AM-11:45AM	278.9
Soil6	T1S4	The Pool in Central Park, Manhattan	1	4	2	0	15-Jun-18	9:30AM-11:45AM	272.4
Soil7	T1S5	The Pool in Central Park, Manhattan	1	5	0	0	15-Jun-18	9:30AM-11:45AM	252
Soil8	T1S5-2	The Pool in Central Park, Manhattan	1	5	0	1	15-Jun-18	9:30AM-11:45AM	263.1
Soil9	T1S6	The Pool in Central Park, Manhattan	1	6	1	0	15-Jun-18	9:30AM-11:45AM	266.8
Soil10	T1S6	The Pool in Central Park, Manhattan	1	6	2	0	15-Jun-18	9:30AM-11:45AM	236.2
Soil11	T1S7	The Pool in Central Park, Manhattan	1	7	0	0	15-Jun-18	9:30AM-11:45AM	256
Soil12	T1S8	The Pool in Central Park, Manhattan	1	8	1	0	15-Jun-18	9:30AM-11:45AM	266.1
Soil13	T1S8	The Pool in Central Park, Manhattan	1	8	2	0	15-Jun-18	9:30AM-11:45AM	244
Soil14	T1S9	The Pool in Central Park, Manhattan	1	9	0	0	15-Jun-18	9:30AM-11:45AM	651.3
Soil15	T1S10	The Pool in Central Park, Manhattan	1	10	1	0	15-Jun-18	9:30AM-11:45AM	318.7
Soil16	T1S10	The Pool in Central Park, Manhattan	1	10	2	0	15-Jun-18	9:30AM-11:45AM	268.7
Soil17	T1S10-2	The Pool in Central Park, Manhattan	1	10	0	1	15-Jun-18	9:30AM-11:45AM	277.5
Soil18	T1S11	The Pool in Central Park, Manhattan	1	11	0	0	15-Jun-18	9:30AM-11:45AM	299.9
Soil19	T1S12	The Pool in Central Park, Manhattan	1	12	1	0	15-Jun-18	9:30AM-11:45AM	282.6
Soil20	T1S12	The Pool in Central Park, Manhattan	1	12	2	0	15-Jun-18	9:30AM-11:45AM	260.4
Soil21	T1S13	The Pool in Central Park, Manhattan	1	13	0	0	15-Jun-18	9:30AM-11:45AM	292.3
Soil22	T1S14	The Pool in Central Park, Manhattan	1	14	1	0	15-Jun-18	9:30AM-11:45AM	360.7
Soil23	T1S14	The Pool in Central Park, Manhattan	1	14	2	0	15-Jun-18	9:30AM-11:45AM	278
Soil24	T1S15	The Pool in Central Park, Manhattan	1	15	0	0	15-Jun-18	9:30AM-11:45AM	341.5
Soil25	T1S15-2	The Pool in Central Park, Manhattan	1	15	0	1	15-Jun-18	9:30AM-11:45AM	319.8
Soil26	T1S16	The Pool in Central Park, Manhattan	1	16	1	0	15-Jun-18	9:30AM-11:45AM	261
Soil27	T1S16	The Pool in Central Park, Manhattan	1	16	2	0	15-Jun-18	9:30AM-11:45AM	267.8
Soil28	T1S17	The Pool in Central Park, Manhattan	1	17	0	0	15-Jun-18	9:30AM-11:45AM	250
Soil29	T1S18	The Pool in Central Park, Manhattan	1	18	1	0	15-Jun-18	9:30AM-11:45AM	281.8
Soil30	T1S18	The Pool in Central Park, Manhattan	1	18	2	0	15-Jun-18	9:30AM-11:45AM	253.3
Soil31	T1S19	The Pool in Central Park, Manhattan	1	19	0	0	15-Jun-18	9:30AM-11:45AM	267.1
Soil32	T1S20	The Pool in Central Park, Manhattan	1	20	1	0	15-Jun-18	9:30AM-11:45AM	329.2
Soil33	T1S20	The Pool in Central Park, Manhattan	1	20	2	0	15-Jun-18	9:30AM-11:45AM	323.4
Soil34	T1S20-2	The Pool in Central Park, Manhattan	1	20	0	1	15-Jun-18	9:30AM-11:45AM	429.3
Soil35	T1S21	The Pool in Central Park, Manhattan	1	21	0	0	15-Jun-18	9:30AM-11:45AM	376.9
Soil36	T1S22	The Pool in Central Park, Manhattan	1	22	1	0	15-Jun-18	9:30AM-11:45AM	259.1

Soil37	T1S22	The Pool in Central Park, Manhattan	1	22	2	0	15-Jun-18	9:30AM-11:45AM	273.5
Soil38	T1S23	The Pool in Central Park, Manhattan	1	23	0	0	15-Jun-18	9:30AM-11:45AM	289.1
Soil39	T1S24	The Pool in Central Park, Manhattan	1	24	1	0	15-Jun-18	9:30AM-11:45AM	329
Soil40	T1S24	The Pool in Central Park, Manhattan	1	24	2	0	15-Jun-18	9:30AM-11:45AM	367.2
Soil41	T1S25	The Pool in Central Park, Manhattan	1	25	0	0	15-Jun-18	9:30AM-11:45AM	370.6
Soil42	T1S26	The Pool in Central Park, Manhattan	1	26	1	0	15-Jun-18	9:30AM-11:45AM	272.4
Soil43	T1S26	The Pool in Central Park, Manhattan	1	26	2	0	15-Jun-18	9:30AM-11:45AM	294.6
Soil44	T1S27	The Pool in Central Park, Manhattan	1	27	0	0	15-Jun-18	9:30AM-11:45AM	347.4
Soil45	T1S28	The Pool in Central Park, Manhattan	1	28	1	0	15-Jun-18	9:30AM-11:45AM	262.7
Soil46	T1S28	The Pool in Central Park, Manhattan	1	28	2	0	15-Jun-18	9:30AM-11:45AM	251.5
Soil47	T1S28-2	The Pool in Central Park, Manhattan	1	28	0	1	15-Jun-18	9:30AM-11:45AM	276.9
Soil48	T2S1	The Pool in Central Park, Manhattan	2	1	1	0	22-Jun-18	9:30AM-11:45AM	327.2
Soil49	T2S1	The Pool in Central Park, Manhattan	2	1	2	0	22-Jun-18	9:30AM-11:45AM	307.3
Soil50	T2S2	The Pool in Central Park, Manhattan	2	2	0	0	22-Jun-18	9:30AM-11:45AM	258.9
Soil51	T2S3	The Pool in Central Park, Manhattan	2	3	1	0	22-Jun-18	9:30AM-11:45AM	263
Soil52	T2S3	The Pool in Central Park, Manhattan	2	3	2	0	22-Jun-18	9:30AM-11:45AM	238.9
Soil53	T2S4	The Pool in Central Park, Manhattan	2	4	0	0	22-Jun-18	9:30AM-11:45AM	263.7
Soil54	T2S5	The Pool in Central Park, Manhattan	2	5	1	0	22-Jun-18	9:30AM-11:45AM	292.1
Soil55	T2S5	The Pool in Central Park, Manhattan	2	5	2	0	22-Jun-18	9:30AM-11:45AM	308
Soil56	T2S5-2	The Pool in Central Park, Manhattan	2	5	0	1	22-Jun-18	9:30AM-11:45AM	293.1
Soil57	T2S6	The Pool in Central Park, Manhattan	2	6	0	0	22-Jun-18	9:30AM-11:45AM	266.6
Soil58	T2S7	The Pool in Central Park, Manhattan	2	7	1	0	22-Jun-18	9:30AM-11:45AM	296.1
Soil59	T2S7	The Pool in Central Park, Manhattan	2	7	2	0	22-Jun-18	9:30AM-11:45AM	274.4
Soil60	T2S8	The Pool in Central Park, Manhattan	2	8	0	0	22-Jun-18	9:30AM-11:45AM	282.2
Soil61	T2S9	The Pool in Central Park, Manhattan	2	9	1	0	22-Jun-18	9:30AM-11:45AM	371.1
Soil62	T2S9	The Pool in Central Park, Manhattan	2	9	2	0	22-Jun-18	9:30AM-11:45AM	347.2
Soil63	T2S10	The Pool in Central Park, Manhattan	2	10	0	0	22-Jun-18	9:30AM-11:45AM	272
Soil64	T2S10-2	The Pool in Central Park, Manhattan	2	10	0	1	22-Jun-18	9:30AM-11:45AM	380.1
Soil65	T2S11	The Pool in Central Park, Manhattan	2	11	1	0	22-Jun-18	9:30AM-11:45AM	400.6
Soil66	T2S11	The Pool in Central Park, Manhattan	2	11	2	0	22-Jun-18	9:30AM-11:45AM	407.8
Soil67	T2S12	The Pool in Central Park, Manhattan	2	12	0	0	22-Jun-18	9:30AM-11:45AM	326.9
Soil68	T2S13	The Pool in Central Park, Manhattan	2	13	1	0	22-Jun-18	9:30AM-11:45AM	476.4
Soil69	T2S13	The Pool in Central Park, Manhattan	2	13	2	0	22-Jun-18	9:30AM-11:45AM	392.1
Soil70	T2S14	The Pool in Central Park, Manhattan	2	14	0	0	22-Jun-18	9:30AM-11:45AM	328.4
Soil71	T2S15	The Pool in Central Park, Manhattan	2	15	1	0	22-Jun-18	9:30AM-11:45AM	315.5
Soil72	T2S15	The Pool in Central Park, Manhattan	2	15	2	0	22-Jun-18	9:30AM-11:45AM	276.6
Soil73	T2S15-2	The Pool in Central Park, Manhattan	2	15	0	1	22-Jun-18	9:30AM-11:45AM	313.8
Soil74	T2S16	The Pool in Central Park, Manhattan	2	16	0	0	22-Jun-18	9:30AM-11:45AM	261.4
Soil75	T2S17	The Pool in Central Park, Manhattan	2	17	1	0	22-Jun-18	9:30AM-11:45AM	226.8

Soil76	T2S17	The Pool in Central Park, Manhattan	2	17	2	0	22-Jun-18	9:30AM-11:45AM	103.9
Soil77	T2S18	The Pool in Central Park, Manhattan	2	18	0	0	22-Jun-18	9:30AM-11:45AM	274.1
Soil78	T2S19	The Pool in Central Park, Manhattan	2	19	1	0	22-Jun-18	9:30AM-11:45AM	251.8
Soil79	T2S19	The Pool in Central Park, Manhattan	2	19	2	0	22-Jun-18	9:30AM-11:45AM	250.2
Soil80	T2S20	The Pool in Central Park, Manhattan	2	20	0	0	22-Jun-18	9:30AM-11:45AM	294.8
Soil81	T2S20-2	The Pool in Central Park, Manhattan	2	20	0	1	22-Jun-18	9:30AM-11:45AM	314.4
Soil82	T2S21	The Pool in Central Park, Manhattan	2	21	1	0	22-Jun-18	9:30AM-11:45AM	342.6
Soil83	T2S21	The Pool in Central Park, Manhattan	2	21	2	0	22-Jun-18	9:30AM-11:45AM	338.8
Soil84	T2S22	The Pool in Central Park, Manhattan	2	22	0	0	22-Jun-18	9:30AM-11:45AM	249.4
Soil85	T2S23	The Pool in Central Park, Manhattan	2	23	1	0	22-Jun-18	9:30AM-11:45AM	294.5
Soil86	T2S23	The Pool in Central Park, Manhattan	2	23	2	0	22-Jun-18	9:30AM-11:45AM	274.4
Soil87	T2S24	The Pool in Central Park, Manhattan	2	24	0	0	22-Jun-18	9:30AM-11:45AM	292.6
Soil88	T2S25	The Pool in Central Park, Manhattan	2	25	1	0	22-Jun-18	9:30AM-11:45AM	359.7
Soil89	T2S25	The Pool in Central Park, Manhattan	2	25	2	0	22-Jun-18	9:30AM-11:45AM	330.8
Soil90	T2S25-2	The Pool in Central Park, Manhattan	2	25	0	1	22-Jun-18	9:30AM-11:45AM	279
Soil91	T2S26	The Pool in Central Park, Manhattan	2	26	0	0	22-Jun-18	9:30AM-11:45AM	249.1
Soil92	T2S27	The Pool in Central Park, Manhattan	2	27	1	0	22-Jun-18	9:30AM-11:45AM	293.1
Soil93	T2S27	The Pool in Central Park, Manhattan	2	27	2	0	22-Jun-18	9:30AM-11:45AM	290.5
Soil94	T2S28	The Pool in Central Park, Manhattan	2	28	0	0	22-Jun-18	9:30AM-11:45AM	271.2
Soil97	T3S1	The Pool in Central Park, Manhattan	3	1	1	0	24-Oct-18	9:30AM-11:45AM	207.3
Soil98	T3S1	The Pool in Central Park, Manhattan	3	1	2	0	24-Oct-18	9:30AM-11:45AM	201.5
Soil99	T3S2	The Pool in Central Park, Manhattan	3	2	0	0	24-Oct-18	9:30AM-11:45AM	183
Soil100	T3S2-2	The Pool in Central Park, Manhattan	3	2	0	1	24-Oct-18	9:30AM-11:45AM	197.2
Soil101	T3S3	The Pool in Central Park, Manhattan	3	3	1	0	24-Oct-18	9:30AM-11:45AM	204.7
Soil102	T3S3	The Pool in Central Park, Manhattan	3	3	2	0	24-Oct-18	9:30AM-11:45AM	231.2
Soil103	T3S4	The Pool in Central Park, Manhattan	3	4	0	0	24-Oct-18	9:30AM-11:45AM	236.2
Soil104	T3S4-2	The Pool in Central Park, Manhattan	3	4	0	1	24-Oct-18	9:30AM-11:45AM	207.1
Soil105	T3S5	The Pool in Central Park, Manhattan	3	5	1	0	24-Oct-18	9:30AM-11:45AM	216.2
Soil106	T3S5	The Pool in Central Park, Manhattan	3	5	2	0	24-Oct-18	9:30AM-11:45AM	221.1
Soil107	T3S6	The Pool in Central Park, Manhattan	3	6	0	0	24-Oct-18	9:30AM-11:45AM	223.3
Soil108	T3S6-2	The Pool in Central Park, Manhattan	3	6	0	1	24-Oct-18	9:30AM-11:45AM	309.3
Soil109	T3S7	The Pool in Central Park, Manhattan	3	7	1	0	24-Oct-18	9:30AM-11:45AM	257.3
Soil110	T3S7	The Pool in Central Park, Manhattan	3	7	2	0	24-Oct-18	9:30AM-11:45AM	251.5
Soil111	T3S8	The Pool in Central Park, Manhattan	3	8	0	0	24-Oct-18	9:30AM-11:45AM	194.2
Soil112	T3S8-2	The Pool in Central Park, Manhattan	3	8	0	1	24-Oct-18	9:30AM-11:45AM	209.6
Soil113	T3S9	The Pool in Central Park, Manhattan	3	9	1	0	24-Oct-18	9:30AM-11:45AM	340.4
Soil114	T3S9	The Pool in Central Park, Manhattan	3	9	2	0	24-Oct-18	9:30AM-11:45AM	389.3
Soil115	T3S10	The Pool in Central Park, Manhattan	3	10	0	0	24-Oct-18	9:30AM-11:45AM	241.6
Soil116	T3S10-2	The Pool in Central Park, Manhattan	3	10	0	1	24-Oct-18	9:30AM-11:45AM	246.8

Soil117	T3S11	The Pool in Central Park, Manhattan	3	11	1	0	24-Oct-18	9:30AM-11:45AM	247.9
Soil118	T3S11	The Pool in Central Park, Manhattan	3	11	2	0	24-Oct-18	9:30AM-11:45AM	276.6
Soil119	T3S12	The Pool in Central Park, Manhattan	3	12	0	0	24-Oct-18	9:30AM-11:45AM	224.2
Soil120	T3S12-2	The Pool in Central Park, Manhattan	3	12	0	1	24-Oct-18	9:30AM-11:45AM	236.4
Soil121	T3S13	The Pool in Central Park, Manhattan	3	13	1	0	24-Oct-18	9:30AM-11:45AM	220.1
Soil122	T3S13	The Pool in Central Park, Manhattan	3	13	2	0	24-Oct-18	9:30AM-11:45AM	288.2
Soil123	T3S14	The Pool in Central Park, Manhattan	3	14	0	0	24-Oct-18	9:30AM-11:45AM	239.5
Soil124	T3S14-2	The Pool in Central Park, Manhattan	3	14	0	1	24-Oct-18	9:30AM-11:45AM	238.7
Soil125	T3S15	The Pool in Central Park, Manhattan	3	15	1	0	24-Oct-18	9:30AM-11:45AM	262
Soil126	T3S15	The Pool in Central Park, Manhattan	3	15	2	0	24-Oct-18	9:30AM-11:45AM	229.8
Soil127	T3S16	The Pool in Central Park, Manhattan	3	16	0	0	24-Oct-18	9:30AM-11:45AM	193.2
Soil128	T3S16-2	The Pool in Central Park, Manhattan	3	16	0	1	24-Oct-18	9:30AM-11:45AM	250.5
Soil129	T3S17	The Pool in Central Park, Manhattan	3	17	1	0	24-Oct-18	9:30AM-11:45AM	189.8
Soil130	T3S17	The Pool in Central Park, Manhattan	3	17	2	0	24-Oct-18	9:30AM-11:45AM	203.9
Soil131	T3S18	The Pool in Central Park, Manhattan	3	18	0	0	24-Oct-18	9:30AM-11:45AM	286
Soil132	T3S18-2	The Pool in Central Park, Manhattan	3	18	0	1	24-Oct-18	9:30AM-11:45AM	264.8
Soil133	T3S19	The Pool in Central Park, Manhattan	3	19	1	0	24-Oct-18	9:30AM-11:45AM	189.5
Soil134	T3S19	The Pool in Central Park, Manhattan	3	19	2	0	24-Oct-18	9:30AM-11:45AM	220.3
Soil135	T3S20	The Pool in Central Park, Manhattan	3	20	0	0	24-Oct-18	9:30AM-11:45AM	253.3
Soil136	T3S20-2	The Pool in Central Park, Manhattan	3	20	0	1	24-Oct-18	9:30AM-11:45AM	258.4
Soil137	T3S21	The Pool in Central Park, Manhattan	3	21	1	0	24-Oct-18	9:30AM-11:45AM	355.2
Soil138	T3S21	The Pool in Central Park, Manhattan	3	21	2	0	24-Oct-18	9:30AM-11:45AM	346.8
Soil139	T3S22	The Pool in Central Park, Manhattan	3	22	0	0	24-Oct-18	9:30AM-11:45AM	244.2
Soil140	T3S22-2	The Pool in Central Park, Manhattan	3	22	0	1	24-Oct-18	9:30AM-11:45AM	269.2
Soil141	T3S23	The Pool in Central Park, Manhattan	3	23	1	0	24-Oct-18	9:30AM-11:45AM	260.7
Soil142	T3S23	The Pool in Central Park, Manhattan	3	23	2	0	24-Oct-18	9:30AM-11:45AM	277.2
Soil143	T3S24	The Pool in Central Park, Manhattan	3	24	0	0	24-Oct-18	9:30AM-11:45AM	273.4
Soil144	T3S24-2	The Pool in Central Park, Manhattan	3	24	0	1	24-Oct-18	9:30AM-11:45AM	363.2
Soil145	T3S25	The Pool in Central Park, Manhattan	3	25	1	0	24-Oct-18	9:30AM-11:45AM	266.9
Soil146	T3S25	The Pool in Central Park, Manhattan	3	25	2	0	24-Oct-18	9:30AM-11:45AM	298
Soil147	T3S26	The Pool in Central Park, Manhattan	3	26	0	0	24-Oct-18	9:30AM-11:45AM	325.4
Soil148	T3S26-2	The Pool in Central Park, Manhattan	3	26	0	1	24-Oct-18	9:30AM-11:45AM	288.4
Soil149	T3S27	The Pool in Central Park, Manhattan	3	27	1	0	24-Oct-18	9:30AM-11:45AM	285.4
Soil150	T3S27	The Pool in Central Park, Manhattan	3	27	2	0	24-Oct-18	9:30AM-11:45AM	322.8
Soil151	T3S28	The Pool in Central Park, Manhattan	3	28	0	0	24-Oct-18	9:30AM-11:45AM	265.5
Soil152	T3S28-2	The Pool in Central Park, Manhattan	3	28	0	1	24-Oct-18	9:30AM-11:45AM	221.1

Supplementary Note: Full description of variance and covariance decomposition models

Contents

1 Variance decomposition model

1.1 Overview

Let X_i be a random variable denoting the abundance of a given bacterial taxon i measured from either 16S rRNA or whole-metagenome shotgun sequencing. Although we focus on single taxon abundances, the following also applies for total bacterial loads in each sample. We let the measured abundances of taxon i collected at different time points of a time series study reflect draws from an underlying distribution $p(X_i)$, which we refer to as the marginal distribution of X_i. We assume that there are three contributions to the total observed variability of X_i. The first corresponds to any set of temporal factors that systematically change from one day to the next to drive changes in taxa abundances over time. This may encompass environmental factors, interspecies interactions and competition, and neutral drift in abundances. The second contribution reflects heterogeneity in the abundance of bacteria across different spatial locations in a given environment. Notably, this spatial sampling variability is inherent to some time series studies such as those of the gut microbiome, where measurements from fecal samples collected at different time points necessarily come from different spatial locations. We take the spatial sampling variability to reflect variation arising from differences in niche size and availability or random dispersal of taxa abundances. The third corresponds to experimental noise associated with bacterial DNA extraction from samples, PCR amplification and sequencing itself. We collectively refer to these experimental sources of variability as technical noise. Our goal is to derive expressions for each of these sources of variability and demonstrate how one may estimate them from experiments.

1.2 Decomposing the variance in microbiota abundances

We first demonstrate how we can use the law of total variance to decompose measured bacterial abundance variances into the three contributions described in the previous section.

1.2.1 Defining the space and time variables

As bacterial abundances change over time, we would expect that abundances of the same taxa i measured at the same time point across different spatial locations in the environment would be more similar to each other than those collected from different time points. However, even at the same time point, measured abundances will not be identical due to both variation across different spatial locations and technical noise. Mathematically, the variability in abundances of taxon i at a fixed time $T = t$ defines a conditional random variable X_i^t with distribution $p(X_i|T = t)$, where T is a time-associated random variable that captures the collective temporal state of the community has underlying distribution $p(T)$. This conditional distribution itself may change when T realizes different values at different time points. Importantly, however, when conditioned on a particular time point $T = t$, the variance of the conditional distribution $Var(X_i^t)$ reflects only spatial heterogeneity and technical noise.

At a given point in time, we may also choose a location from which to collect a sample to sequence. We can therefore define another random variable $X_i^{t,s}$ with probability distribution $p(X_i|T = t, S = s)$ representing the abundance of taxon i measured from this fixed time point $T = t$ and spatial location $S = s$. Here, S is a space-associated random variable with distribution $p(S|T = t)$ that at a given point in time, changes with the particular spatial location from which taxon i is collected and measured. Conditioning on both space and time, we have eliminated any biological sources of variability and the variance of taxon i, $Var(X_i^{t,s})$, simply reflects technical noise. The distributions $p(X_i)$, $p(X_i|T = t)$, $p(X_i|T = t, S = s)$ and their hierarchical relationships are illustrated in Fig. S1, with the fecal microbiome shown as the model ecosystem.

1.2.2 Variance decomposition

Using the law of total variance, we now decompose the total abundance variance of X_i into components associated with time, spatial sampling location and technical noise. In the following, E and Var denote the expectation and variance of a random variable respectively, and subscripts denote the underlying distribution ($p(T)$ or $p(S|T)$) with respect to which the operation is performed. Beginning with the definition of the variance of X_i,

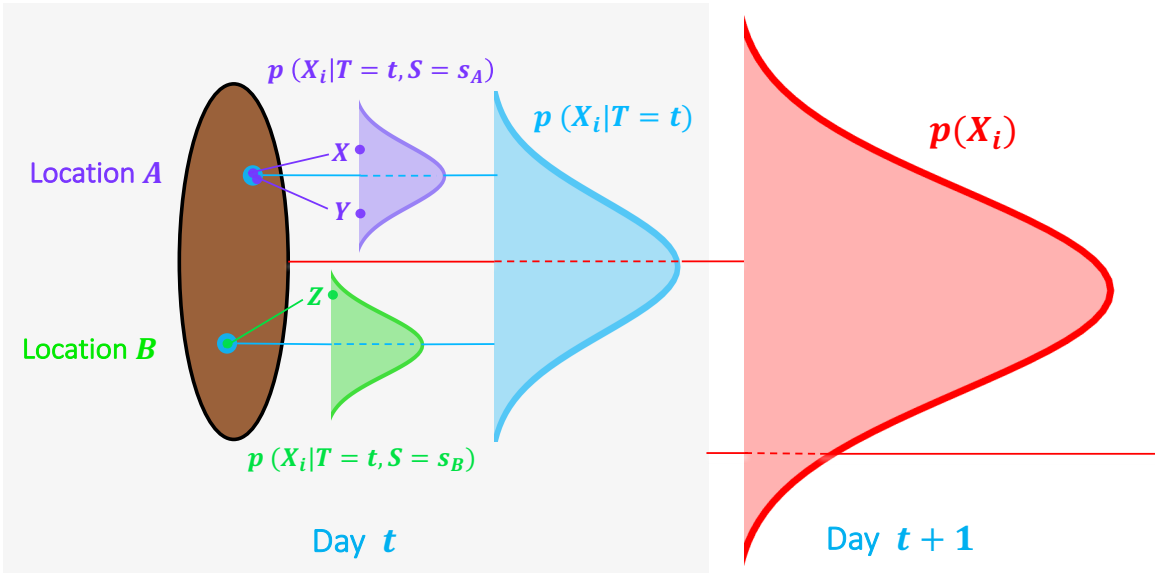


Fig. S1: Statistical model for the decomposition of bacterial abundance variances in the fecal microbiome.

$$
\begin{aligned}
Var(X_i) &= E(X_i^2) - [E(X_i)]^2 \\
&= E_T E_{S|T} E(X_i^2|S,T) - [E_T E_{S|T} E(X_i|S,T)]^2 \\
&= E_T E_{S|T} Var(X_i|S,T) + E_T E_{S|T}[E(X_i|S,T)]^2 - [E_T E_{S|T} E(X_i|S,T)]^2 \\
&= E_T E_{S|T} Var(X_i|S,T) + E_T Var_{S|T} E(X_i|S,T) + E_T[E_{S|T} E(X_i|S,T|)]^2 - [E_T E_{S|T} E(X_i|S,T)]^2 \qquad (1)
\end{aligned}
$$

Thus,

$$
Var(X_i) = \underbrace{E_T E_{S|T} Var(X_i|S,T)}_{\text{Technical } (\langle \sigma_N^2 \rangle_{S,T})} + \underbrace{E_T Var_{S|T} E(X_i|S,T)}_{\text{Spatial sampling } (\langle \sigma_S^2 \rangle_T)} + \underbrace{Var_T E_{S|T} E(X_i|S,T)}_{\text{Temporal } (\sigma_T^2)} \qquad (2)
$$

The terms in the last line correspond to the contributions of technical, spatial sampling, and temporal factors to the total variance of X_i, which we denote with the symbols $\langle \sigma_N^2 \rangle_{S,T}$, $\langle \sigma_S^2 \rangle_T$ and σ_T^2 respectively. Equation (2) is the law of total variance generalized to multiple conditional random variables. Indeed, the right-most term can be recognized as the variance of X_i explained by the time random variable T. The second term reflects the spatial sampling variance of taxa abundances conditioned on time, then averaged over time. The first term is simply the technical variability conditioned on a spatial location and time point, and then jointly averaged over space and time.

1.3 Model-driven experimental approach

We now describe the approach we use to estimate the three contributions to total abundance variability derived in Section 1.2. We show mathematically how our model can be used to estimate each of these terms with only a minimal set of experiments. Notably, the approach we adopt is conceptually similar to the dual reporter method originally described by Elowtiz et. al. [1] used to separate intrinsic versus extrinsic sources of noise in the gene expression profiles of single cells [2, 3, 4].

1.3.1 Experimental setup

Extending the notation described above, we denote X_i, Y_i and Z_i to be random variables representing abundances of a taxon i made from three separate measurements at each point in time from the community. Let us assume that the abundances X_i and Y_i are made from the same exact location (by sequencing the sample twice), whereas Z_i is measured from an independent location. As X_i, Y_i and Z_i correspond to the same bacterial taxon, their marginal distributions are equivalent. Importantly, however, because the abundances X_i,Y_i and Z_i are sampled together at the same time points, there exists a covariance structure driven by shared but unknown temporal factors tending to cause their abundances to collectively increase or decrease from one day to the next. Therefore, although X_i, Y_i and Z_i are identically distributed, they are not independent. In addition, by letting X_i and Y_i correspond to abundances measured not only at the same time point but also from the same spatial location, the covariance between X_i and Y_i is also driven by shared spatial factors that result in similar taxa abundances across different locations.

We can break the described covariance structures by conditioning abundances on time and spatial location. Specifically, when conditioning abundances on time $T = t$, the pairs X_i^t, Z_i^t and Y_i^t, Z_i^t have become sets of independent draws from the distribution $p(X_i|T = t)$. Experimentally, independence is achieved if Z_i^t is sampled from a location in the community independent from X_i^t and Y_i^t. Note that while X_i^t and Y_i^t have the same underlying distribution $p(X_i|T = t)$, they are not independent as their values covary across space. However, further conditioning of X_i^t and Y_i^t on spatial location results in the conditional random variables $X_i^{t,s}$ and $Y_i^{t,s}$ which are indeed independent draws from the distribution $p(X_i|T = t, S = s)$. The assumption of independence is reasonable, as $X_i^{t,s}$ and $Y_i^{t,s}$ are simply technical replicates.

Therefore, our replicate sampling protocol goes as follows: at each time point, we make three abundance measurements for all bacterial taxa $i \in 1..N$. Two of these abundance measurements (X_i and Y_i) are made from the same spatial location in a given environment. The third (Z_i) is measured from a separate, independent location. From an experimental standpoint, X_i and Y_i correspond to technical replicates while X_i,Z_i and Y_i,Z_i correspond to spatial replicates.

1.4 Derivation of statistical estimators for variance decomposition

We can now derive the statistical estimators for each of the terms in equation (2) using the complete hierarchical model described in Section 1.3. We begin by showing that under the specified model, the first two moments of the marginal distributions of X_i, Y_i, and Z_i are indeed identical.

Mean: Using the law of total expectation,

$$E(X_i) = E_T E_{S|T} E(X_i|S,T) = E_T E_{S|T} E(Y_i|S,T) = E(Y_i) \tag{3}$$

$$E(X_i) = E_T E(X_i|T) = E_T E(Z_i|T) = E(Z_i) \tag{4}$$

Variance: By the law of total variance,

$$\begin{aligned} Var(X_i) &= E_T E_{S|T} Var(X_i|S,T) + E_T Var_{S|T} E(X_i|S,T) + Var_T E_{S|T} E(X_i|S,T) \\ &= E_T E_{S|T} Var(Y_i|S,T) + E_T Var_{S|T} E(Y_i|S,T) + Var_T E_{S|T} E(Y_i|S,T) = Var(Y_i) \end{aligned} \tag{5}$$

$$Var(X_i) = E_T Var(X_i|T) + Var_T E(X_i|T) = E_T Var(Z_i|T) + Var_T E(Z_i|T) = Var(Z_i) \tag{6}$$

We have now laid the groundwork for the following derivations of statistical estimators for each of the terms in equation (2). This is the primary result of the variance decomposition model.

1.4.1 Variance associated with time

$$\sigma_T^2 = Var_T E_{S|T} E(X_i|S,T) = Var_T E(X_i|T)$$

$$= E_T[E(X_i|T)]^2 - [E_T E(X_i|T)]^2$$

$$= E_T[E(X_i|T)E(Z_i|T)] - E_T E(X_i|T) E_T E(Z_i|T)$$

$$= E_T E(X_i Z_i|T) - E_T E(X_i|T) E_T E(Z_i|T)$$

$$= E(X_i Z_i) - E(X_i)E(Z_i) = Cov(X_i, Z_i) \tag{7}$$

Hence, the time-associated variability is simply the covariance between X_i and Z_i. Its unbiased estimator is:

$$\hat{\sigma_T^2} = \frac{1}{n-1} \sum_{t=1}^{n} (x_i^t - \bar{x_i})(z_i^t - \bar{z_i}) \tag{8}$$

1.4.2 Variance associated with spatial sampling location

$$\langle \sigma_S^2 \rangle_T = E_T Var_{S|T} E(X_i|S,T) = E_T E_{S|T}[E(X_i|S,T)]^2 - E_T[E_{S|T} E(X_i|S,T)]^2$$

$$= E_T E_{S|T}[E(X_i|S,T)]^2 - E_T[E(X_i|T)]^2$$

$$= E_T E_{S|T}[E(X_i|S,T)E(Y_i|S,T)] - E_T[E(Z_i|T)][E(Y_i|T)]$$

$$= E_T E_{S|T} E(X_i Y_i|S,T) - E_T E(Z_i Y_i|T)$$

$$= E(X_i Y_i) - E(Z_i Y_i)$$

$$= E(X_i Y_i - Z_i Y_i) - E(X_i)E(Y_i) + E(Z_i)E(Y_i)$$

$$= E((X_i - Z_i)Y_i) - E(X_i - Z_i)E(Y_i) = Cov(X_i - Z_i, Y_i) \tag{9}$$

Similar to time, the spatial sampling-associated variance reduces to the covariance between $X_i - Z_i$ and Y_i and is estimated by:

$$\hat{\langle \sigma_S^2 \rangle_T} = \frac{1}{n-1} \sum_{t=1}^{n} [(x_i^t - z_i^t) - (\bar{x_i} - \bar{z_i})](y_i^t - \bar{y_i}) \tag{10}$$

1.4.3 Technical noise

$$\langle \sigma_N^2 \rangle_{S,T} = E_T E_{S|T} Var(X_i|S,T) = E_T E_{S|T} E(X_i^2|S,T) - E_T E_{S|T}[E(X_i|S,T)]^2$$

$$= \frac{1}{2}[E_T E_{S|T} E(X_i^2|S,T) - 2E_T E_{S|T}[E(X_i|S,T)E(Y_i|S,T)] + E_T E_{S|T} E(Y_i^2|S,T)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E_T E_{S|T} E(X_i Y_i|S,T) + E(Y_i^2)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E(X_i Y_i) + E(Y_i^2)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E(X_i Y_i) + E(Y_i^2) - E(X_i)^2 + 2E(X_i)E(Y_i) - E(Y_i)^2]$$

$$= \frac{1}{2}[E([X_i - Y_i]^2) - [E(X_i - Y_i)]^2]$$

$$= \frac{1}{2} Var(X_i - Y_i) \tag{11}$$

Finally, the technical variability is simply the variance of the difference between X_i and Y_i, whose unbiased estimator is:

$$\langle \hat{\sigma_N^2} \rangle_{S,T} = \frac{1}{2(n-1)} \sum_{t=1}^{n} [(x_i^t - y_i^t) - (\bar{x_i} - \bar{y_i})]^2 \tag{12}$$

We note that the protocol and derivations presented above represent the minimum number of sample collections and preparations required for proper variance decomposition in the data. However, the above derivations also apply if multiple (more than two) spatial or technical replicates are taken at each time point. In this case, we may average over pairwise combinations of spatial or technical replicates in the data (i.e. average over different realizations of X, Y, and Z in the data). Therefore, the presented framework may be applied to a study design involving an arbitrary number of spatial or technical replicates, so long as these numbers are greater than or equal to two. Finally, we note that if (1) the number of statistical samples is small and (2) the variances are small, the variance estimators defined above may return negative or NaN values. In our analysis, we omit these estimations.

1.5 Generalizing the hierarchy

In the sections above, we have implicitly imposed a hierarchy in our model. Namely, quantities in each term of equation (2) are first averaged with respect to $p(X_i|S,T)$, followed by $p(S|T)$ and finally $p(T)$. One can also imagine reversing this hierarchy; that is, averaging with respect to $p(T|S)$ second, followed by $p(S)$ last. Experimentally, this would correspond to drawing temporal replicates from a fixed spatial location, then averaging quantities over various locations. In ecosystems such as the human gut microbiome, the hierarchy described in earlier sections arises naturally, as $p(S|T)$ and $p(T)$ are experimentally accessible from fecal samples, while $p(T|S)$ and $p(S)$ are not. In other words, it is only possible to draw spatial replicates from a fixed time point, then average bacterial abundance measurements over multiple time points. It is not possible to measure abundances from the same spatial location on fecal samples obtained from two different days. However, in other microbial communities, such as those found in the soil, one can imagine an alternative hierarchical sampling protocol in which different spatial locations in the ecosystem are each sampled on two separate days. Equation (2) may then be used to decompose bacterial abundance variances using this inverted hierarchy:

$$Var(X_i) = \underbrace{E_S E_{T|S} Var(X_i|S,T)}_{\text{Technical } (\langle \sigma_N^2 \rangle_{S,T})} + \underbrace{E_S Var_{T|S} E(X_i|S,T)}_{\text{Temporal } (\langle \sigma_T^2 \rangle_S)} + \underbrace{Var_S E_{T|S} E(X_i|S,T)}_{\text{Spatial sampling } (\sigma_S^2)} \quad (13)$$

The first term reflecting technical noise remains the same. The second term now reflects a temporal variance that is averaged over sampling locations. The third reflects the variance explained by spatial sampling location. Indeed, we used this inverted hierarchy to carry out sampling of a soil microbial community in Central Park. We also note that the spatial sampling variability here may be defined rather loosely. For example, if one were to collect two human microbiome samples at different time points from a large cohort of individuals and sequence one of these two samples twice, the spatial sampling variability could be referred to as inter-individual variability while the temporal variability would correspond to intra-individual variability, with the noise term remaining the same.

Finally, we note that abundance measurements may be made from multiple different spatial locations at multiple different time points, with multiple sequencing replicates performed at each time and location. Although one may arrive at the different variance contributions directly by following the prescription of equation (2) and using the typical estimators for mean and variance, this experimental protocol quickly becomes prohibitive for longer time series studies. Importantly, our hierarchical replicate sampling protocol makes such data collection unnecessary.

2 Covariance decomposition model

2.1 Overview

We now extend our variance decomposition model for single taxa to a generalized covariance decomposition for all pairs of taxa. We let X_i and X_j denote the abundances of taxa i and j measured together from the same sample (i.e. sequenced from the same spatial location in a given environment). As with the case of single taxon variances, we assume that the total abundance covariance between taxa i and j (across different samples collected over time) may be attributed to underlying temporal, spatial and technical sources. Intuitively, the temporal contribution results from the covariation in overall abundances (averaged over all spatial locations in the community) of taxa i and j from one day to the next. In addition, however, species abundances may be correlated across different spatial locations at a given time point, which is captured by the spatial contribution to the total covariance. Finally, technical factors may result in correlated noise, potentially arising from sources such as similar DNA extraction efficiencies or primer and amplification biases. Again, our goal is to derive expressions for each of these covariance sources and demonstrate how one may estimate them experimentally using the same protocol described in Section 1.

2.2 Decomposing the covariance in pairs of bacterial species abundances

2.2.1 Total and conditional joint distributions

The total abundance covariance of taxa i and j may be calculated from the simple experiment: draw a single sample from a random spatial location in the environment at each time point and sequence the abundances of i and j. Mathematically, we may consider the bivariate random variable $\vec{X}$ with components comprising the measured abundances (X_i, X_j) and define a total joint distribution $p(X_i, X_j)$. We refer to this as the total joint distribution because in the data collection process, we have marginalized over time, space and technical noise. In contrast, one can imagine a second experiment where at a given time point, multiple samples are obtained across various spatial locations in the community. This defines another distribution, the conditional joint distribution $p(X_i, X_j|T = t)$ for some fixed time point t, where variances and covariances now reflect both underlying spatial factors as well as technical noise. Finally, by fixing both time and sampling location $S = s$ and re-extracting and sequencing bacterial DNA multiple times, a third distribution $p(X_i, X_j|T = t, S = s)$ can be defined. Here, variances and covariances reflect purely technical sources. The hierarchical relationships of these distributions are illustrated in Fig. S2. We show in the next section how the total covariance $Cov(X_i, X_j)$ between taxa i and j may be decomposed by making use of the described conditional joint distributions.

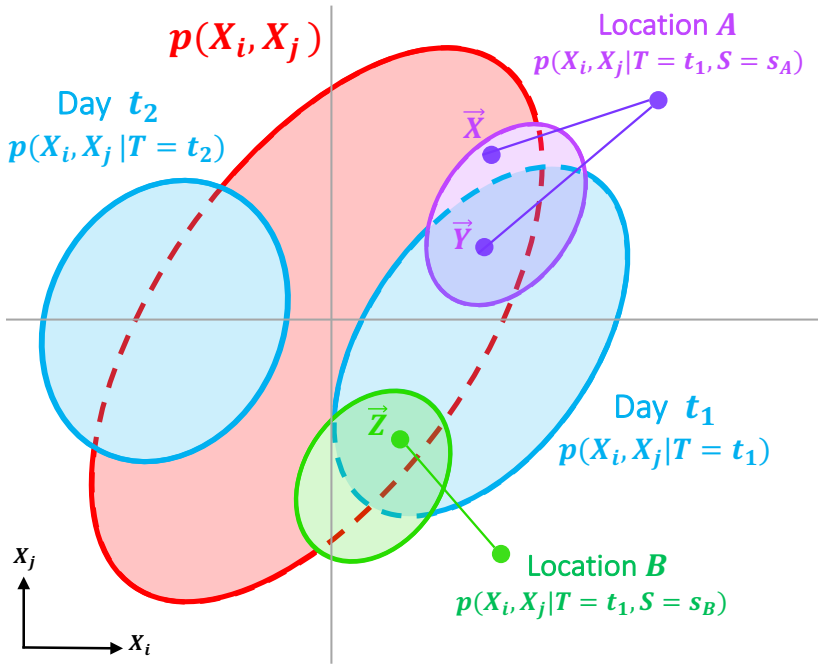


Fig. S2: Statistical model for the decomposition of bacterial abundance covariances.

2.2.2 *Covariance decomposition*

The total covariance corresponding to the distribution $p(X_i, X_j)$ can be written as:

$$
\begin{aligned}
Cov(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\
&= E_T E_{S|T} E(X_i X_j|S,T) - E(X_i)E(X_j) \\
&= E_T E_{S|T} Cov(X_i, X_j|S,T) + E_T E_{S|T}[E(X_i|S,T)E(X_j|S,T)] - E(X_i)E(X_j) \\
&= E_T E_{S|T} Cov(X_i, X_j|S,T) + E_T Cov_{S|T}(E(X_i|S,T), E(X_j|S,T)) \\
&+ E_T[E_{S|T}E(X_i|S,T)E_{S|T}E(X_j|S,T)] - E(X_i)E(X_j)
\end{aligned} \tag{14}
$$

Mirroring the variance decomposition, we arrive at:

$$
Cov(X_i, X_j) = \underbrace{E_T E_{S|T} Cov(X_i, X_j|S,T)}_{\text{Technical } (\langle \sigma^i \sigma^j{}_N \rangle_{S,T})} + \underbrace{E_T Cov_{S|T}(E(X_i|S,T), E(X_j|S,T))}_{\text{Spatial sampling } (\langle \sigma^i \sigma^j{}_S \rangle_T)} + \underbrace{Cov_T(E(X_i|T), E(X_j|T))}_{\text{Temporal } (\sigma^i \sigma^j{}_T)} \tag{15}
$$

Note that we can obtain correlations by simply dividing each term in equation (15) by marginal standard deviations.

2.3 Derivation of statistical estimators for covariance decomposition

Let X_i, Y_i and Z_i and X_j, Y_j and Z_j denote abundances of taxa i and j respectively as described in Section 1.3.1. That is, the pairs (X_i, Z_i) and (Y_i, Z_i) correspond to spatial replicates while (X_i, Y_i) denote technical replicates. As before, when conditioning on time $T = t$, the bivariate random variable pairs $\vec{X}^t = (X_i^t, X_j^t)$ and $\vec{Z}^t = (Z_i^t, Z_j^t)$, and $\vec{Y}^t = (Y_i^t, Y_j^t)$ and $\vec{Z}^t = (Z_i^t, Z_j^t)$ correspond to two i.i.d. draws from the same underlying distribution $p(X_i, X_j|T = t)$, with a covariance matrix structured by both spatial and technical sources. We will now define two additional and equivalent conditional joint distributions $p(X_i, Z_j|T = t)$, and $p(Y_i, Z_j|T = t)$ where the abundance of taxon i from one spatial replicate X_i^t or Y_i^t is paired with the abundance of taxon j from the second spatial replicate Z_j^t. Note here that while the conditional marginal distributions of $p(X_i, Z_j|T = t)$ and $p(Y_i, Z_j|T = t)$ are identical to those of the distribution $p(X_i, X_j|T = t)$ as demonstrated in Section 1.3, the random variables X_i^t and Z_j^t, and Y_i^t and Z_j^t are independent of one another when conditioned on time, as abundances from each pair come from different spatial locations and sequencing realizations. Along similar lines, the bivariate random variables $\vec{X}^{t,s} = (X_i^{t,s}, X_j^{t,s})$ and $\vec{Y}^{t,s} = (Y_i^{t,s}, Y_j^{t,s})$ correspond to random variables drawn from the distribution $p(X_i, X_j|T = t, S = s)$, where correlations between i and j are driven purely by technical sources. Again, we may preserve conditional marginal distributions while eliminating covariances by defining the distribution $p(X_i, Y_j|T = t, S = s)$. With this in mind, we will now derive statistical estimators for each of the terms in equation (15).

2.3.1 Covariance associated with time

$$
\begin{aligned}
\sigma^i\sigma^j{}_T &= Cov_T(E(X_i|T), E(X_j|T)) \\
&= E_T[E(X_i|T)E(X_j|T)] - E_T E(X_i|T) E_T E(X_j|T) \\
&= E_T[E(X_i|T)E(Z_j|T)] - E_T E(X_i|T) E_T E(Z_j|T) \\
&= E_T E(X_i Z_j|T) - E(X_i)E(Z_j) \\
&= Cov(X_i, Z_j)
\end{aligned} \tag{16}
$$

Analogous to the variance decomposition, the unbiased estimator for the time-associated covariance is:

$$
\hat{\sigma^i}\hat{\sigma^j}{}_T = \frac{1}{n-1}\sum_{t=1}^{n}(x_i^t - \bar{x_i})(z_j^t - \bar{z_j}) \tag{17}
$$

2.3.2 Covariance associated with spatial sampling location

$$\langle \sigma^i \sigma^j{}_S \rangle_T = E_T Cov_{S|T}(E(X_i|S,T), E(X_j|S,T))$$

$$= E_T E_{S|T}[E(X_i|S,T)E(X_j|S,T)] - E_T[E_{S|T}E(X_i|S,T)E_{S|T}E(X_j|S,T)]$$

$$= E_T E_{S|T}[E(X_i|S,T)E(X_j|S,T)] - E_T[E(X_i|T)E(X_j|T)]$$

$$= E_T E_{S|T}[E(X_i|S,T)E(Y_j|S,T)] - E_T[E(Z_i|T)E(Y_j|T)]$$

$$= E_T E_{S|T}E(X_i Y_j|S,T) - E_T E(Z_i Y_j|T)$$

$$= E(X_i Y_j) - E(Z_i Y_j)$$

$$= E(X_i Y_j) - E(Z_i Y_j) - E(X_i)E(Y_j) + E(Z_i)E(Y_j)$$

$$= E((X_i - Z_j)Y_j) + E(X_i - Z_i)E(Y_j)$$

$$= Cov(X_i - Z_i, Y_j) \tag{18}$$

The space-associated covariance is given by:

$$\langle \sigma^i \sigma^j{}_S \rangle_T = \frac{1}{n-1} \sum_{t=1}^{n} [(x_i^t - z_i^t) - (\bar{x_i} - \bar{z_i})](y_j^t - \bar{y_j}) \tag{19}$$

2.3.3 Covariance associated with technical noise

$$\langle \sigma^i \sigma^j{}_N \rangle_{S,T} = E_T E_{S|T} Cov(X_i, X_j | S, T)$$

$$= E_T E_{S|T} E(X_i X_j | S, T) - E_T E_{S|T} [E(X_i | S, T) E(X_j | S, T)]$$

$$= \frac{1}{2} [E_T E_{S|T} E(X_i X_j | S, T) - E_T E_{S|T} [E(X_i | S, T) E(Y_j | S, T)]$$
$$- E_T E_{S|T} [E(Y_i | S, T) E(X_j | S, T)] + E_T E_{S|T} E(Y_i Y_j | S, T)]$$

$$= \frac{1}{2} [E_T E_{S|T} E(X_i X_j | S, T) - E_T E_{S|T} E(X_i Y_j | S, T)$$
$$- E_T E_{S|T} E(Y_i X_j | S, T) + E_T E_{S|T} E(Y_i Y_j | S, T)]$$

$$= \frac{1}{2} [E(X_i X_j) - E(X_i Y_j) - E(Y_i X_j) + E(Y_i Y_j)]$$

$$= \frac{1}{2} [E(X_i X_j - X_i Y_j - Y_i X_j + Y_i Y_j) - E(X_i) E(X_j) + E(X_i) E(Y_j) - E(Y_i) E(X_j) + E(Y_i) E(Y_j)]$$

$$= \frac{1}{2} [E(X_i - Y_i)(X_j - Y_j) - E(X_i - Y_i) E(X_j - Y_j)]$$

$$= \frac{1}{2} Cov(X_i - Y_i, X_j - Y_j) \tag{20}$$

Finally, the covariance associated with technical sources is estimated as:

$$\langle \sigma^i \sigma^j{}_N \rangle_{S,T} = \frac{1}{2(n-1)} \sum_{t=1}^{n} [(x_i^t - y_i^t) - (\bar{x_i} - \bar{y_i})][(x_j^t - y_j^t) - (\bar{x_j} - \bar{y_j})] \tag{21}$$

2.4 Covariances induced by conversion to absolute abundances

Finally, we demonstrate that conversion from relative to absolute abundances typically results in higher covariances for taxa pairs measured in absolute abundances. These higher covariances stem from the variance of total bacterial densities measured from sample to sample that may induce correlations in taxa pairs whose relative abundances are otherwise uncorrelated or even negatively correlated. Let R_i and R_j denote the relative abundances of taxa i and j measured from a single sample. Let us denote A to be the total bacterial abundance density (in units of DNA copies per mg of environmental sample matter). Note that the absolute abundances X_i and X_j are simply calculated as AR_i and AR_j respectively. We will assume that the relative abundances R_i and R_j are independent of A. Then, we may write:

$$
\begin{aligned}
Cov(AR_i, AR_j) &= E(A^2 R_i R_j) - E(A)^2 E(R_i) E(R_j) \\
&= E(A^2) E(R_i R_j) - E(A)^2 E(R_i) E(R_j) \\
&= E(A^2)[Cov(R_i, R_j) + E(R_i) E(R_j)] - E(A)^2 E(R_i) E(R_j) \\
&= E(A^2) Cov(R_i, R_j) + E(R_i) E(R_j)[E(A^2) - E(A)^2)] \\
&= E(A^2) Cov(R_i, R_j) + E(R_i) E(R_j) Var(A) \qquad (22)
\end{aligned}
$$

3 Two component variance and covariance decomposition models

Up to this point, we have considered the general case of three contributions to the variances and covariances of bacterial taxa. It may be useful in some cases to perform a two component variance/covariance decomposition (e.g. to separate any technical from non-technical/biological variability in microbiome sequencing studies). Here we present the mathematical formulation and corresponding statistical estimators for the two component approach.

3.1 Two component variance decomposition

If we are interested in separating biological from technical variability in microbiome studies, we may consider the following two component variance decomposition model:

$$
\begin{aligned}
Var(X_i) &= E(X_i^2) - [E(X_i)]^2 \\
&= E_B E(X_i^2 | B) - [E_B E(X_i | B)]^2 \\
&= E_B Var(X_i | B) + E_B [E(X_i | B)]^2 - [E_B E(X_i | B)]^2 \\
&= \underbrace{E_B Var(X_i | B)}_{\text{Technical } (\langle \sigma_N^2 \rangle_B)} + \underbrace{Var_B E(X_i | B)}_{\text{Biological } (\sigma_B^2)} \qquad (23)
\end{aligned}
$$

Equation (23) is the law of total variance, where the left term reflects the same average technical noise contributing to the total variance of X_i. The right term now reflects the variance of X_i explained by both temporal and spatial factors, where B is a random variable capturing the collective temporal and spatial state of the community. Because, the left term corresponds to noise in the data, the right term captures all the true biological variability of taxon X_i. We may estimate each of these terms as follows:

3.1.1 Biological variability

$$\sigma_B^2 = Var_B E(X_i|B)$$

$$= E_B[E(X_i|B)]^2 - [E_B E(X_i|B)]^2$$

$$= E_B[E(X_i|B)E(Y_i|B)] - E_B E(X_i|B) E_B E(Y_i|B)$$

$$= E_B E(X_i Y_i|B) - E_B E(X_i|B) E_B E(Y_i|B)$$

$$= E(X_i Y_i) - E(X_i)E(Y_i) = Cov(X_i, Y_i) \tag{24}$$

Keeping the same nomenclature as before, the biological variability is the covariance between the technical replicate measurements X_i and Y_i, whose unbiased estimator is:

$$\hat{\sigma_B^2} = \frac{1}{n-1} \sum_{t=1}^{n} (x_i^t - \bar{x_i})(y_i^t - \bar{y_i}) \tag{25}$$

3.1.2 Technical noise

$$\langle \sigma_N^2 \rangle_B = E_B Var(X_i|B) = E_B E(X_i^2|B) - E_B[E(X_i|B)]^2$$

$$= \frac{1}{2}[E_B E(X_i^2|B) - 2E_B[E(X_i|B)E(Y_i|B)] + E_B E(Y_i^2|B)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E_B E(X_i Y_i|B) + E(Y_i^2)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E(X_i Y_i) + E(Y_i^2)]$$

$$= \frac{1}{2}[E(X_i^2) - 2E(X_i Y_i) + E(Y_i^2) - E(X_i)^2 + 2E(X_i)E(Y_i) - E(Y_i)^2]$$

$$= \frac{1}{2}[E([X_i - Y_i]^2) - [E(X_i - Y_i)]^2]$$

$$= \frac{1}{2} Var(X_i - Y_i) \tag{26}$$

The technical variability remains identical to that for the three component variance decomposition:

$$\langle \hat{\sigma_N^2} \rangle_B = \frac{1}{2(n-1)} \sum_{t=1}^{n} [(x_i^t - y_i^t) - (\bar{x_i} - \bar{y_i})]^2 \tag{27}$$

3.2 Generalizing the interpretation of the two component variance decomposition model

While we refer to the biological variability in equation (23) as variability associated with both temporal and spatial factors, this interpretation will depend from study to study. For example, for studies conducted at a single time point, in which multiple spatial sites are sampled across an environment, the biological variability in

equation (23) will simply reflect spatial heterogeneity. The corresponding study design in this scenario would be to collect samples across different spatial locations, and sequence each of these samples twice to obtain X_i and Y_i. For studies interested in investigating population-wide variability of the human microbiome, two technical replicates may be sequenced from single samples collected across individuals. Here, the biological variability will reflect inter-individual variability. Notably, two technical replicates are required from every sample for proper separation of the technical from biological variance contributions.

3.3 Two component covariance decomposition

The total covariance corresponding to the distribution $p(X_i, X_j)$ can be written as:

$$
\begin{aligned}
Cov(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\
&= E_B E(X_i X_j|B) - E(X_i)E(X_j) \\
&= E_B Cov(X_i, X_j|B) + E_B[E(X_i|B)E(X_j|B)] - E(X_i)E(X_j) \\
&= \underbrace{E_B Cov(X_i, X_j|B)}_{\text{Technical } (\langle \sigma^i \sigma^j{}_N \rangle_B)} + \underbrace{Cov_B(E(X_i|B), E(X_j|B))}_{\text{Biological } (\sigma^i \sigma^j{}_B)} \qquad (28)
\end{aligned}
$$

(29)

These terms are estimated in the following sections.

3.3.1 *Biological covariance*

$$
\begin{aligned}
\sigma^i \sigma^j{}_B &= Cov_B(E(X_i|B), E(X_j|B)) \\
&= E_B[E(X_i|B)E(X_j|B)] - E_B E(X_i|B) E_B E(X_j|B) \\
&= E_B[E(X_i|B)E(Y_j|B)] - E_B E(X_i|B) E_B E(Y_j|B) \\
&= E_B E(X_i Y_j|B) - E(X_i)E(Y_j) \\
&= Cov(X_i, Y_j) \qquad (30)
\end{aligned}
$$

Therefore,

$$
\hat{\sigma^i}\hat{\sigma^j}{}_B = \frac{1}{n-1} \sum_{t=1}^{n} (x_i^t - \bar{x_i})(y_j^t - \bar{y_j}) \qquad (31)
$$

3.3.2 Covariance associated with technical noise

$$\langle \sigma^i \sigma^j{}_N \rangle_B = E_B Cov(X_i, X_j|B)$$

$$= E_B E(X_i X_j|B) - E_B[E(X_i|B)E(X_j|B)]$$

$$= \frac{1}{2}[E_B E(X_i X_j|B) - E_B[E(X_i|B)E(Y_j|B)]$$
$$- E_B[E(Y_i|B)E(X_j|B)] + E_B E(Y_i Y_j|B)]$$

$$= \frac{1}{2}[E_B E(X_i X_j|B) - E_B E(X_i Y_j|B)$$
$$- E_B E(Y_i X_j|B) + E_T E(Y_i Y_j|B)]$$

$$= \frac{1}{2}[E(X_i X_j) - E(X_i Y_j) - E(Y_i X_j) + E(Y_i Y_j)]$$

$$= \frac{1}{2}[E(X_i X_j - X_i Y_j - Y_i X_j + Y_i Y_j) - E(X_i)E(X_j) + E(X_i)E(Y_j) - E(Y_i)E(X_j) + E(Y_i)E(Y_j)]$$

$$= \frac{1}{2}[E(X_i - Y_i)(X_j - Y_j) - E(X_i - Y_i)E(X_j - Y_j)]$$

$$= \frac{1}{2}Cov(X_i - Y_i, X_j - Y_j) \tag{32}$$

The covariance associated with technical sources is estimated as:

$$\langle \sigma^i \sigma^j{}_N \rangle_B = \frac{1}{2(n-1)} \sum_{t=1}^{n} [(x_i^t - y_i^t) - (\bar{x_i} - \bar{y_i})][(x_j^t - y_j^t) - (\bar{x_j} - \bar{y_j})] \tag{33}$$

References

[1] M. B. Elowitz, A. J. Levine, and E. D. Siggia. Stochastic Gene Expression in a Single Cell. *Science (80-.).*, 297(August):1183–1187, 2002.

[2] A. Q. Fu and L. Pachter. Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Stat. Appl. Genet. Mol. Biol.*, 15(6):447–471, 2016.

[3] A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proc. Natl. Acad. Sci.*, 108(29):12167–12172, 2011.

[4] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20), 2002.