Relating the metatranscriptome and metagenome of the human gut
SUPPORTING INFORMATION APPENDIX

Eric A. Franzosa[1,2], Xochitl C. Morgan[1,2], Nicola Segata[1], Levi Waldron[1], Joshua Reyes[1], Ashlee M. Earl[2], Georgia Giannoukos[2], Matthew Boylan[3], Dawn Ciulla[2], Dirk Gevers[2], Jacques Izard[4,5], Wendy S. Garrett,[2,6,7], Andrew T. Chan[3,8], Curtis Huttenhower[1,2,*]

*Corresponding author:
chuttenh@hsph.harvard.edu
(617) 432-4912

[1.] Biostatistics Department, Harvard School of Public Health
677 Huntington Avenue, Boston, MA 02115, USA

[2.] The Broad Institute
7 Cambridge Center, Cambridge, MA 02142, USA

[3.] Division of Gastroenterology, Massachusetts General Hospital
55 Fruit Street, GRJ 825C, Boston, MA 02114, USA

[4.] Department of Microbiology, The Forsyth Institute
245 First St, Cambridge, MA 02142, USA

[5.] Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine
188 Longwood Ave, Boston, MA 02115, USA

[6.] Department of Immunology and Infectious Diseases, Harvard School of Public Health
677 Huntington Avenue, Boston, MA 02115, USA

[7.] Department of Medical Oncology, Dana-Farber Cancer Institute
450 Brookline Avenue, Boston, MA 02215, USA

[8.] Channing Division of Network Medicine, Brigham and Women's Hospital
181 Longwood Avenue, Boston, MA, 02115

Supporting Methods

Pan-genome mapping analysis. In addition to the MetaPhlAn- and HUMAnN-based taxonomic and functional profiling described in the main Methods (1, 2), we performed mappings of all read data to reference genomes using a custom approach. Using the MetaPhlAn-based taxonomic profiles of our samples, we identified all species present with relative abundance $\geq 10^{-5}$ (0.001%) in at least one sample. For each of these species, we created a reference pan-genome from all sequenced isolates of that species available in the IMG database (3). Specifically, this involved concatenating all annotated coding sequences from these isolates as a single FASTA file, sorting the file by decreasing sequence length, and then clustering the coding sequences at 95% nucleotide identity using UCLUST (4). This procedure produces a new FASTA file of "seed" sequences, each representing a cluster of genes sharing strong sequence homology. These gene clusters constitute the species' pan-genome. We then mapped quality-trimmed, host- and length-filtered DNA and RNA reads from our samples against the aggregated seed sequences of all pan-genomes using bowtie 2 (5), saving any best hit discovered for each read under the "sensitive" mode. Mapping rates were similar across samples: 31% ± 6% (mean ± std. dev.). These rates are similar to those reported by the Human Microbiome Project (HMP) in its reference genome-based mapping analyses (6). Results of this process are reported for the 8 control stool metagenomes and metatranscriptomes in Supporting Dataset S6.

Evaluating the effects of 24 hours of on-ice incubation on stool meta'omes. Stool samples were collected for this study following an established protocol (7) involving delivery of sample material from subjects to the laboratory on ice within 24 hours. Analyses from the main manuscript demonstrated that, relative to this protocol, two-day simulated shipping of ethanol- or RNAlater-fixed stool aliquots had only a minimal perturbative effect on the gut metagenome and metatranscriptome.

We performed additional experiments to evaluate the effects of the initial on-ice transport relative to an idealized protocol in which stool was immediately flash-frozen upon evacuation. Canine stool was used as a model system for the evaluation. Immediately following sample evacuation, 3 grams of stool were collected in a sterile container and homogenized using the scoop from a feces tube. This sample was then divided using a Puritan DNA-free swab into three control aliquots and three test aliquots stored in 2 mL pyrogen-free tubes. The three control aliquots were flash-frozen in liquid nitrogen within 10 minutes of sample evacuation. The flash frozen aliquots were transported to the laboratory on dry ice within 1 hour. 1.5 mL of 100% ethanol were added to one of the control aliquots, and then all three control aliquots were stored at -80°C. The three test aliquots were placed in an ice chest filled with frozen cold packs, topped with ice, and then allowed to rest at ambient temperature (20°C) for 24 hours to simulate the effects of on-ice transport from subjects to the laboratory. Following this incubation, 1.5 mL of 100% ethanol were added to one of the test aliquots, and then all three test aliquots were stored at -80°C.

Two control aliquots and two test aliquots (including the ethanol-treated aliquots) were extracted and sequenced; the remaining control aliquot and test aliquot were kept in reserve. DNA was extracted from 100 mg of each aliquot using the MoBio PowerLyzer kit (following manufacturer's instructions) and then converted to a DNA library using the Illumina Nextera XT kit. RNA was extracted from 100 mg of each aliquot using the Qiagen AllPrep kit following an established protocol (8). Strand-specific RNA-seq libraries were then created following a custom protocol. Briefly, sample RNA was fragmented and 3' end-tagged with an RNA oligonucleotide

containing a barcode and partial 5' Illumina adapter. The four uniquely tagged RNA pools were then combined and carried through (i) rRNA depletion using the Epicentre Ribo-Zero™ Magnetic Kit (Bacteria), followed by (ii) cDNA synthesis, (iii) ligation to a second oligonucleotide containing a partial Illumina 3' adapter, and finally (iv) amplification with full-length, barcoded Illumina adapter primers. Equimolar amounts of each DNA library and each RNA library were then pooled and sequenced using an Illumina MiSeq instrument, yielding >27 million 101 bp paired-end reads.

DNA and RNA reads were de-multiplexed and then quality- and host-filtered following the procedures outlined in the Methods section of the main text, resulting in a final pool of >24 million reads (2-4 million reads per sample). The *Canis lupus familiaris* genome (GenBank Assembly ID GCA_000002285.2) was used as a basis for identification and removal of host reads. Filtered reads were then profiled with MetaPhlAn (1) to determine taxonomic composition and HUMAnN (2) to determine KEGG pathway abundance. Meta'omic profiles of the flash-frozen control and ice-incubated (test) aliquots are compared in Figure S1.

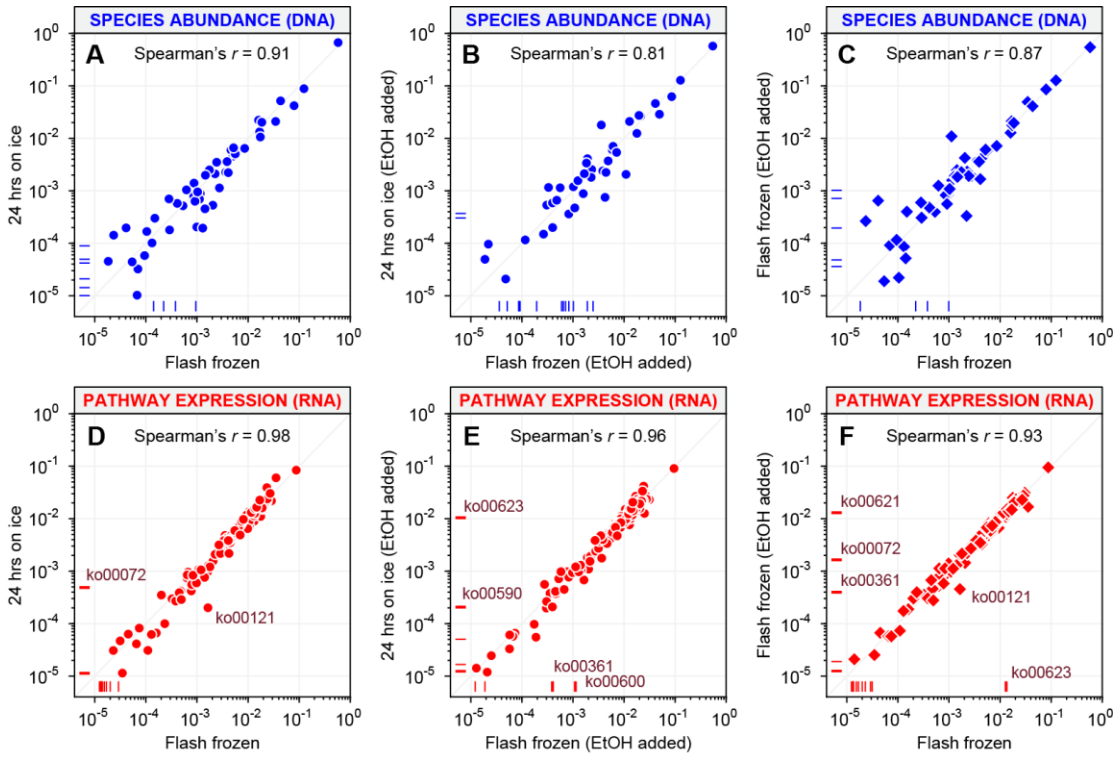
Supporting Tables

Table S1. RIN scores for metatranscriptomic samples subjected to different handling methods. Average scores for all sample handling methods were consistent with a pattern of partial RNA degradation, with no samples classified as strongly degraded (RIN≤3).

Subject	Frozen	EtOH	RNAlater
X310763260	6.2	7.5	6.8
X311245214	5.6	7.9	8.2
X316192082	5.9	6.2	8.4
X316701492	6.5	6.5	8.7
X317690558	5.9	8.5	8.3
X317802115	4.3	4.8	7.5
X317822438	7.2	7.5	6.8
X319146421	4.0	7.3	8.0
Min	4.0	4.8	6.8
Average	**5.7**	**7.0**	**7.8**
Max	7.2	8.5	8.7

Supporting Figures

FIGURE S1. Evaluating the effects of 24 hours of on-ice incubation on stool meta'omes. Aliquots of canine stool were flash frozen or allowed to incubate on ice for 24 hours prior to extraction. Ethanol (EtOH) was added to one aliquot of each type. **(A)** and **(B)** compare the resulting taxonomic (species) profiles of the ice-incubated versus flash-frozen aliquots, with the EtOH-treated aliquots shown in (B). Taxonomic profiles were very similar, suggesting that microbial growth during the on-ice incubation only minimally perturbed the original sample composition. **(C)** The degree of variability between the flash-frozen and ice-incubated aliquots was similar to that observed between the two flash-frozen aliquots, indicating that other stochastic factors (e.g. initial compositional variation among the aliquots) contributed to the variation in (A) and (B). **(D)**, **(E)**, and **(F)** illustrate an analogous series of comparisons at the level of KEGG pathway expression. Pathway expression was well conserved between the ice-incubated and flash-frozen samples, with stochastic sampling effects again playing a role in explaining the observed differences. For example, comparing panels (D) and (F) reveals that ketone metabolism, although it appeared to be up-regulated in the ice-incubated sample from (D), was more likely under-sampled in the flash-frozen sample from (D), as this pathway was clearly expressed in the second (EtOH-treated) flash-frozen sample, as seen in (F). Based on these findings, we conclude that the initial transport of samples from subjects to the laboratory on-ice was not likely to induce large perturbations in the samples' metagenomic and metatranscriptomic composition. (X and Y values reflect relative abundance; zero values are shown as hash marks in the axis margins.)



ko00072: Synthesis and degradation of ketone bodies
ko00121: Secondary bile acid biosynthesis
ko00361: Chlorocyclohexane and chlorobenzene degradation
ko00590: Arachidonic acid metabolism
ko00600: Sphingolipid metabolism
ko00621: Dioxin degradation
ko00623: Toluene degradation

FIGURE S2. Principal Coordinates Analysis (PCoA) of the effect of sample handling method on meta'omic measurements. For species (left column), samples cluster strongly by subject, while sample handling methods are well-mixed. Gene and transcript profiles also tend to cluster by subject (center and right columns). One EtOH-fixed sample (subject #5) is an outlier in the gene profile comparison, but not in the transcript comparison.

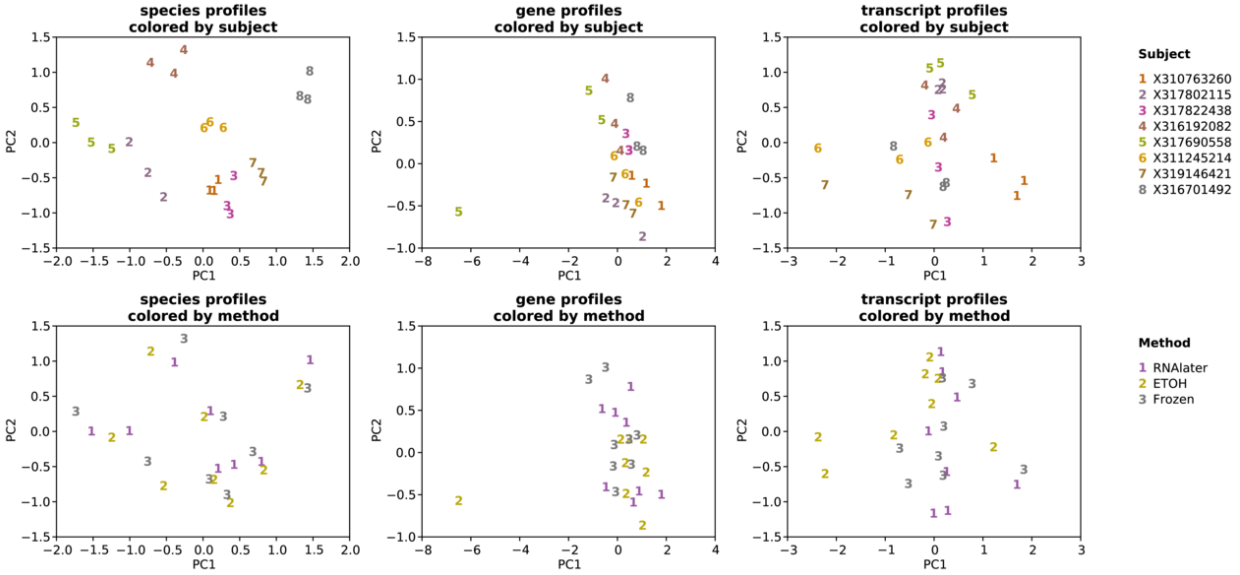


FIGURE S3. Principal Coordinates Analysis (PCoA) of cohort effect (HPFS versus HMP). (TOP) HPFS stool samples are not separated from the HMP stool samples. **(BOTTOM)** HPFS saliva samples cluster between two different oral body sites assessed in the HMP cohort.

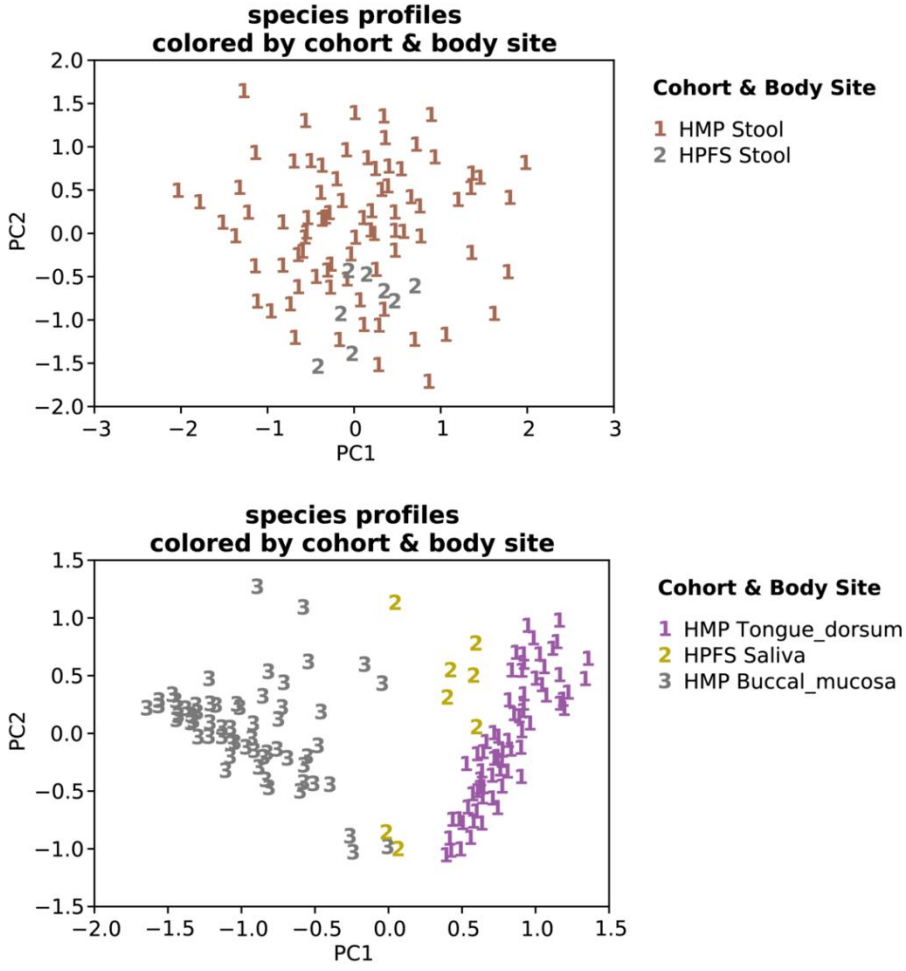


FIGURE S4. Species-specific marker analysis for *Haemophilus parainfluenzae*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the top and bottom panels.

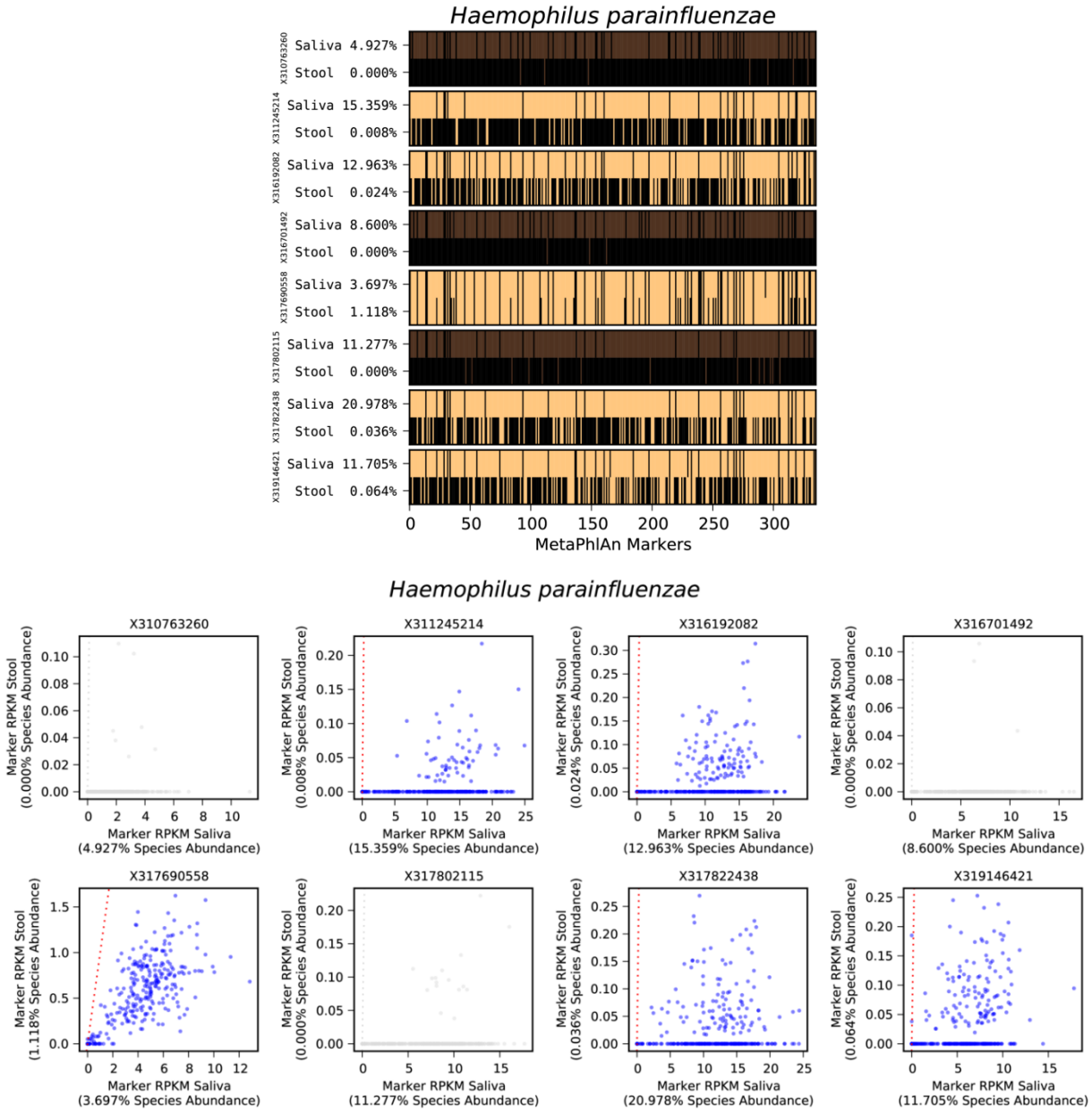


FIGURE S5. Species-specific marker analysis for *Streptococcus salivarius*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the top and bottom panels.

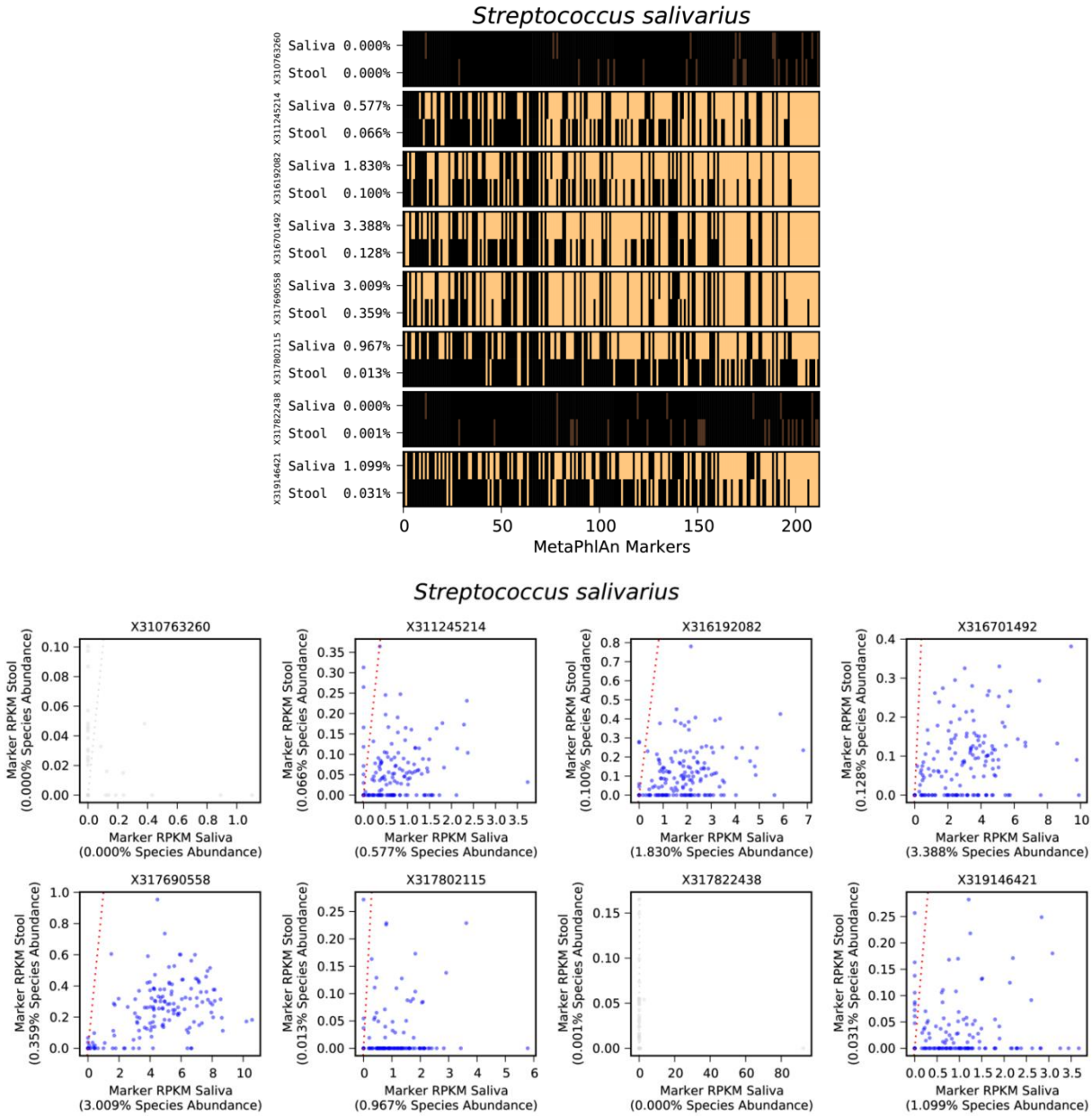


FIGURE S6. Species-specific marker analysis for *Veillonella atypica*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance $\geq 10^{-5}$) are desaturated in both the top and bottom panels.

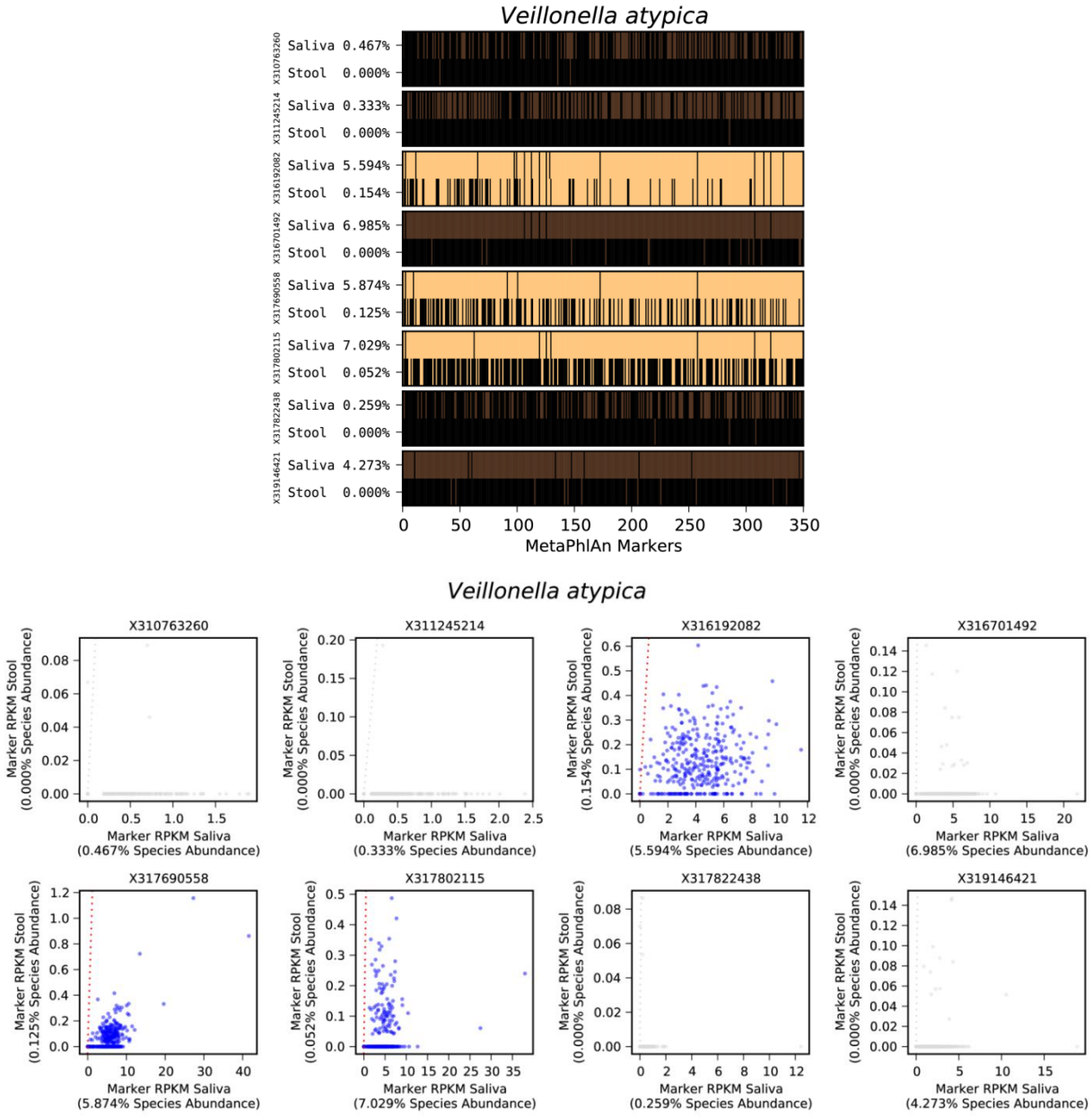


FIGURE S7. Species-specific marker analysis for *Veillonella parvula*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance $\geq 10^{-5}$) are desaturated in both the top and bottom panels.

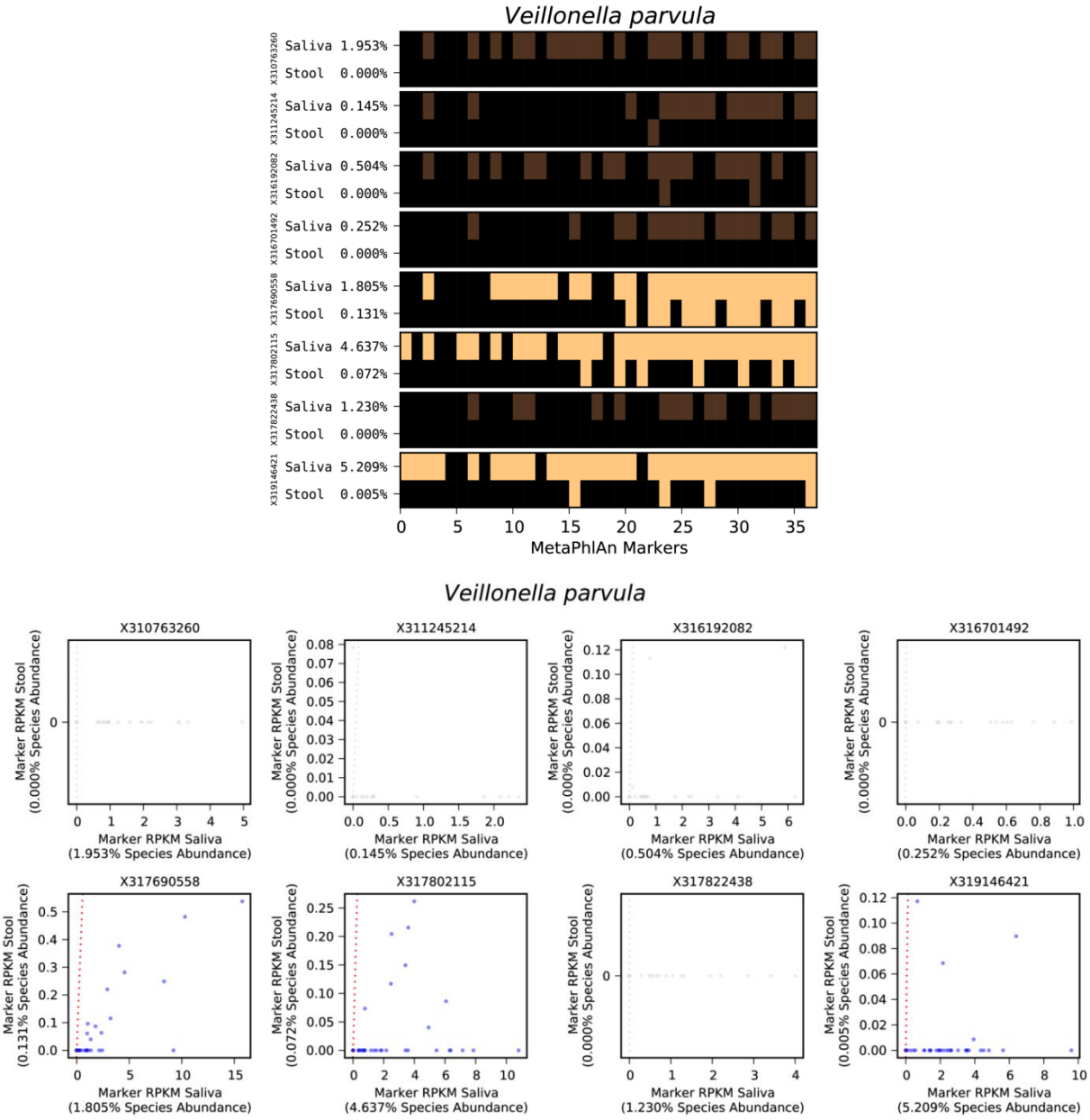


FIGURE S8. Species-specific marker analysis for *Streptococcus parasanguinis*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the top and bottom panels.

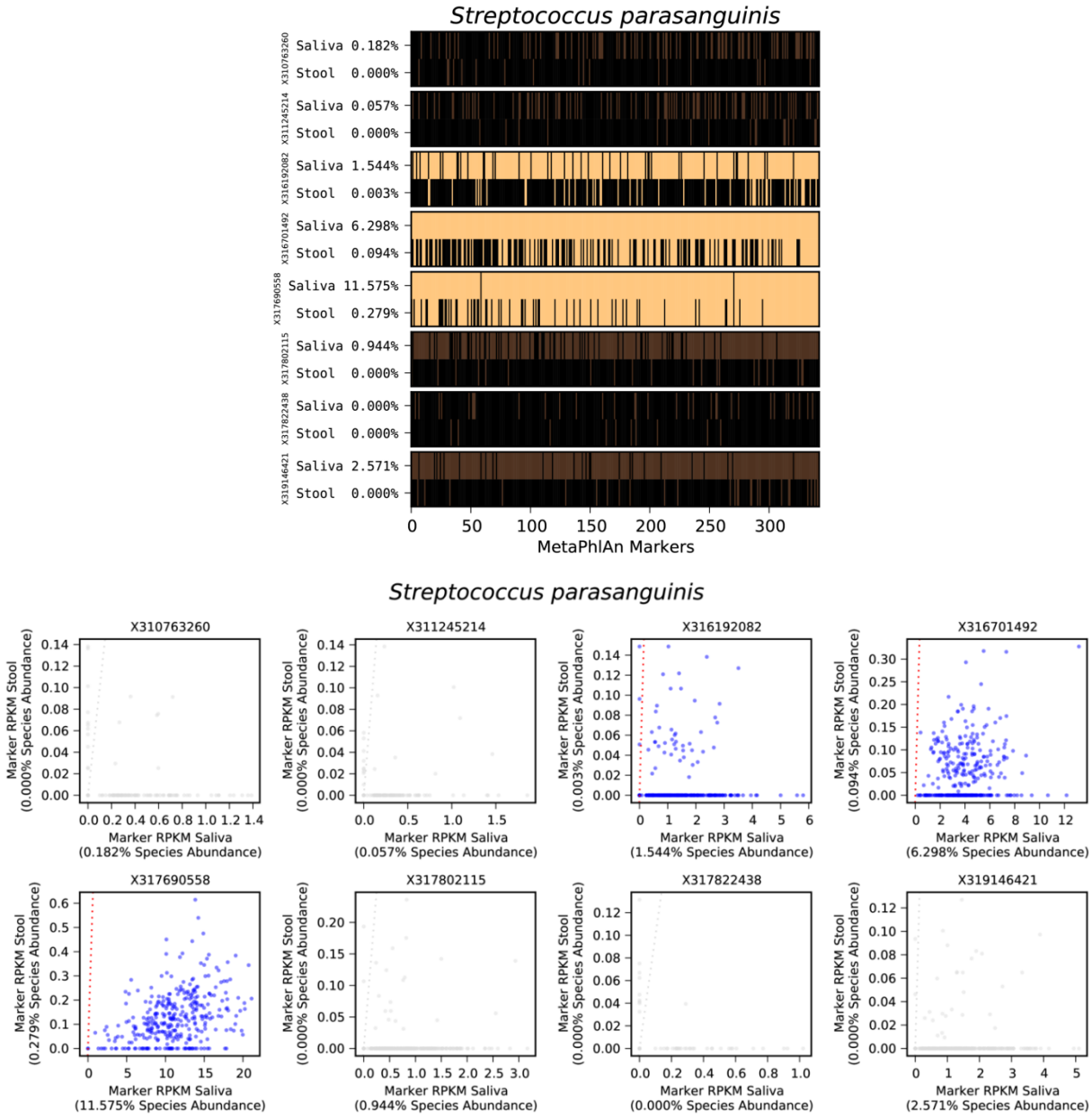


FIGURE S9. Species-specific marker analysis for *Streptococcus australis*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the top and bottom panels.

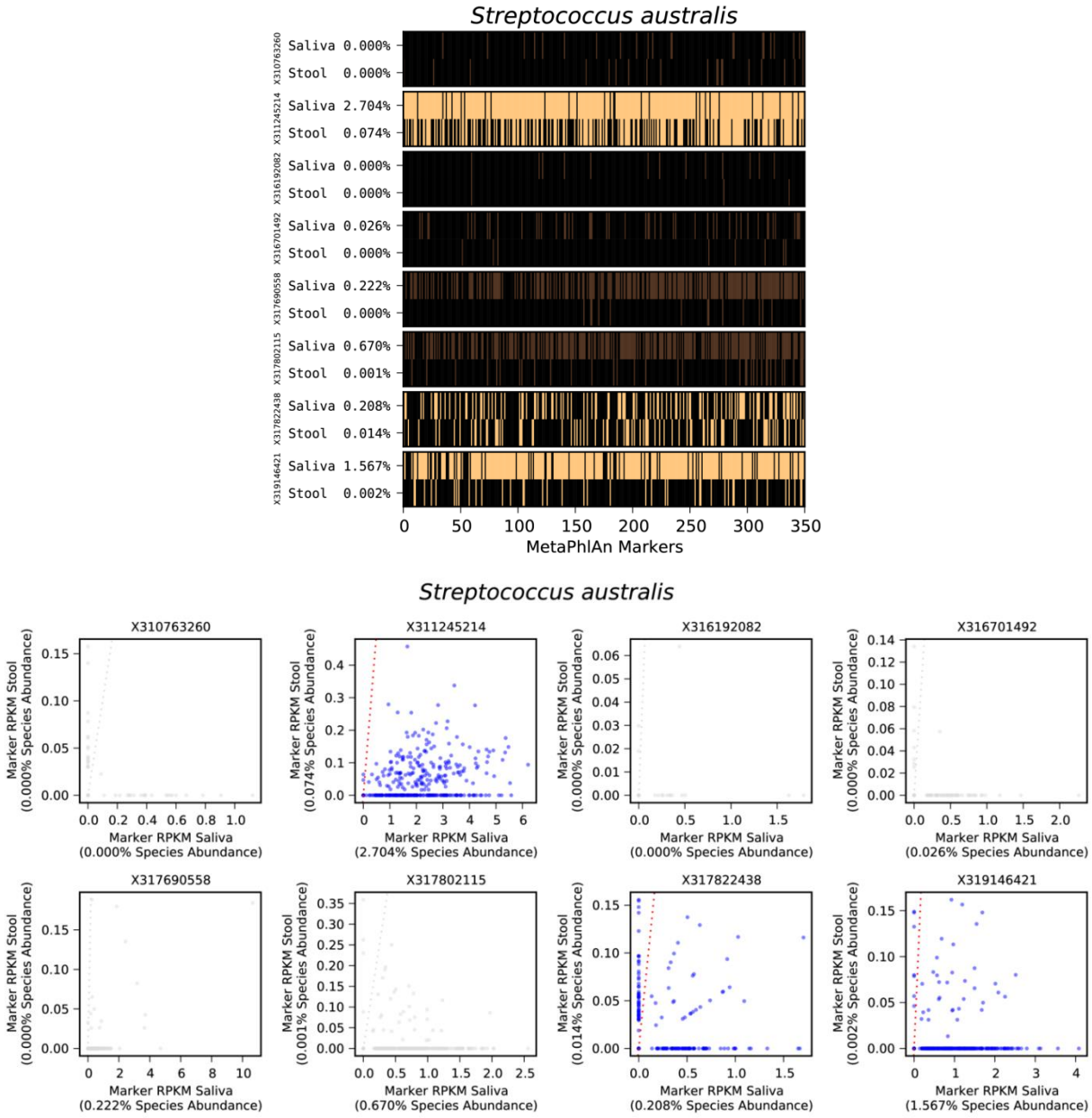


FIGURE S10. Species-specific marker analysis for *Actinomyces odontolyticus*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the top and bottom panels.

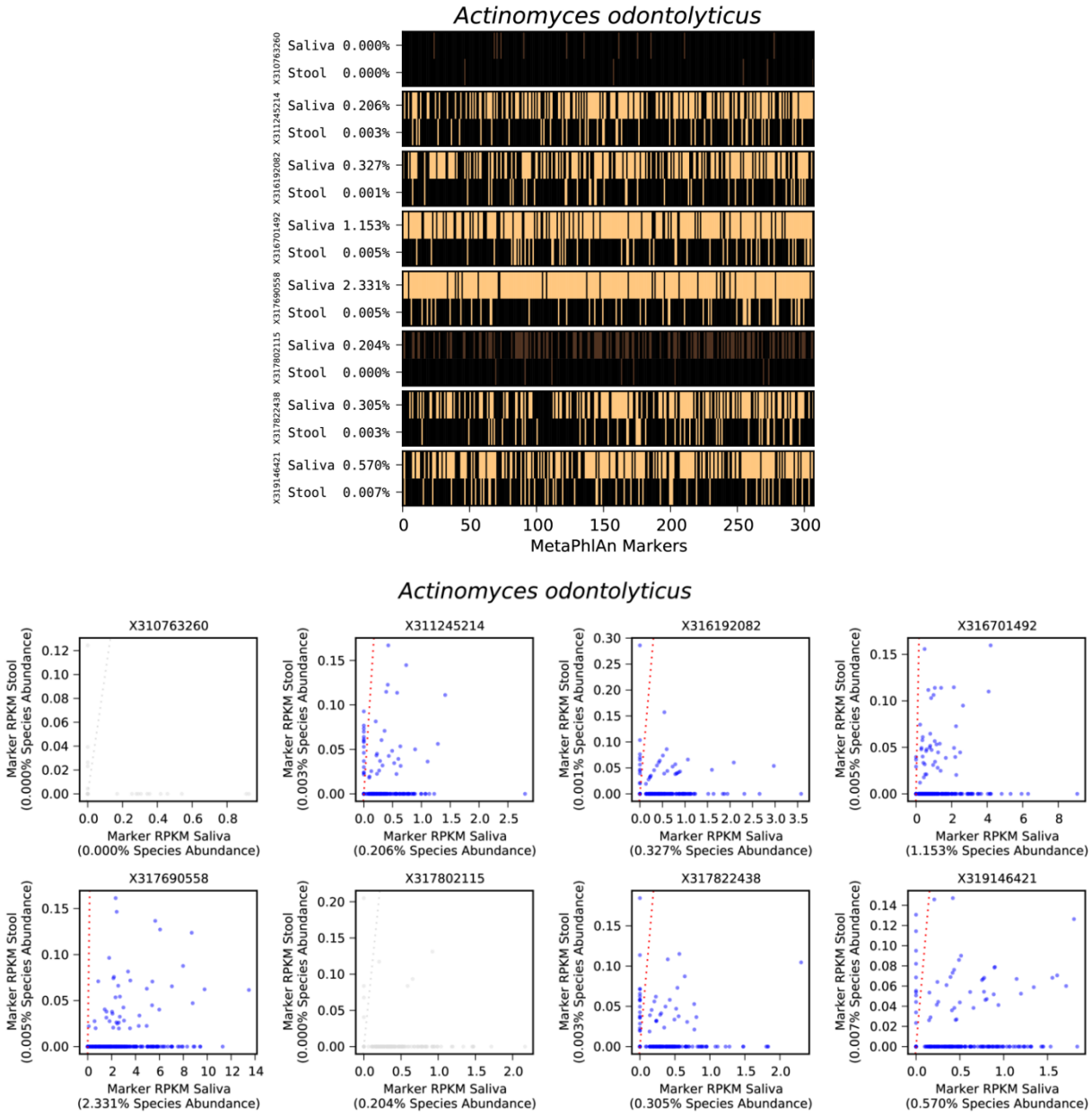


FIGURE S11. Species-specific marker analysis for *Streptococcus sanguinis*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. We rarely observe markers that are exclusively detected in the gut, consistent with a single strain occurring at both body sites within a given subject. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which the species was not confidently detected at both body sites (relative abundance ≥10^{-5}) are desaturated in both the TOP and BOTTOM panels.

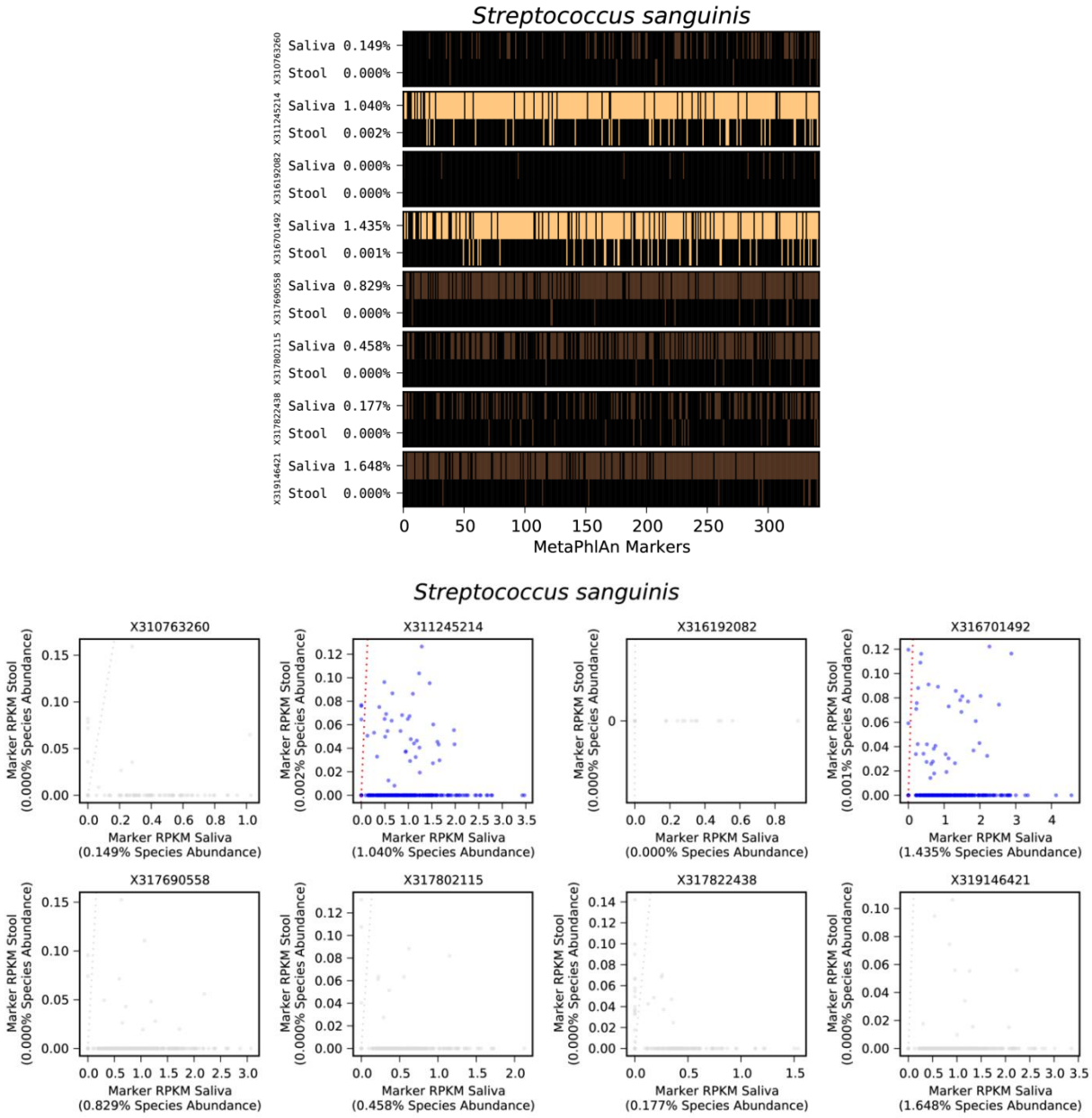


FIGURE S12. Species-specific marker analysis for *Dialister invisus*. (TOP) MetaPhlAn marker barcodes. Marker genes are arranged in genomic order for each sample, with detected markers highlighted in orange. Saliva and stool samples are paired from each individual, with total species abundance shown at the left. Note patterns of between-subject strain variation and within-subject strain conservation. **(BOTTOM)** Quantitative comparison of marker abundance (in RPKM units) at the oral and gut body sites. *Dialister invisus* is an unusual example of a species that co-occurs in subjects' saliva and stool samples but is more abundant in the stool. We rarely observe markers that are exclusively detected in the saliva, consistent with a single strain occurring at both body sites. The dotted red line indicates equal marker abundance in the oral and gut sites. Samples for which *Dialister invisus* was not confidently detected at both body sites (relative abundance $\geq 10^{-5}$) are desaturated in both the top and bottom panels.

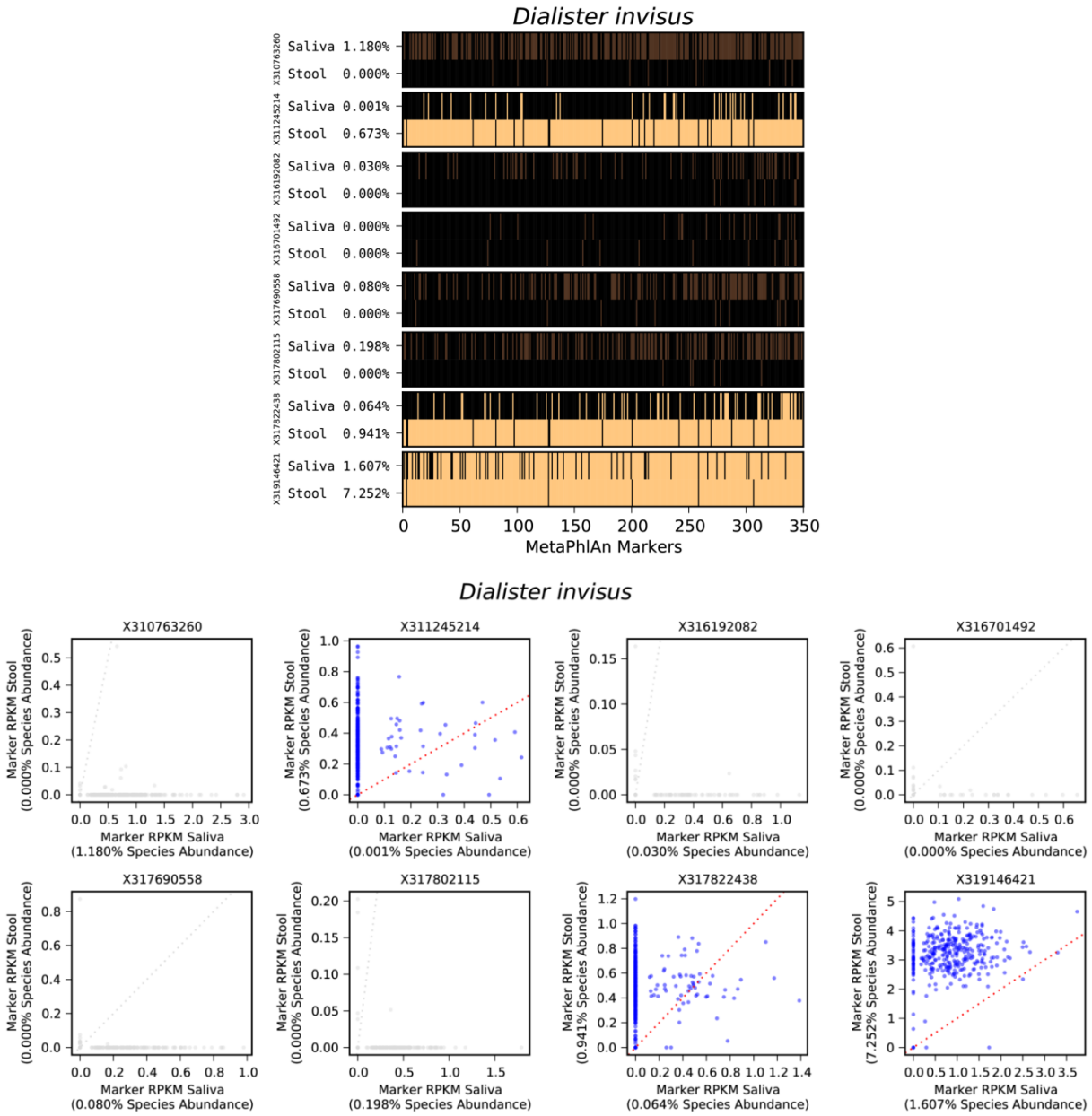


FIGURE S13. Pan-genome recruitment plots. Following the pan-genome mapping procedure described in the Supporting Methods, we converted counts of DNA and RNA read hits to RPKM units and plotted these values for several species-of-interest (*Dialister invisus*, *Alistipes putredinis*, and *Methanobrevibacter smithii*). For each species, only the three samples in which the species was the most abundant at the DNA level are shown. Percentages reflect the coverage of the pan-genome at the gene (blue) and transcript (red) levels. Transcript coverage of the *Dialister invisus* pangenome is unusually low relative to its high coverage at the gene level. *Alistipes putredinis* has strong coverage at both the gene and transcript levels, although its average transcript abundance tends to fall below that of the species' genes. Conversely, *Methanobrevibacter smithii* transcripts are consistently more abundant than their corresponding genes.

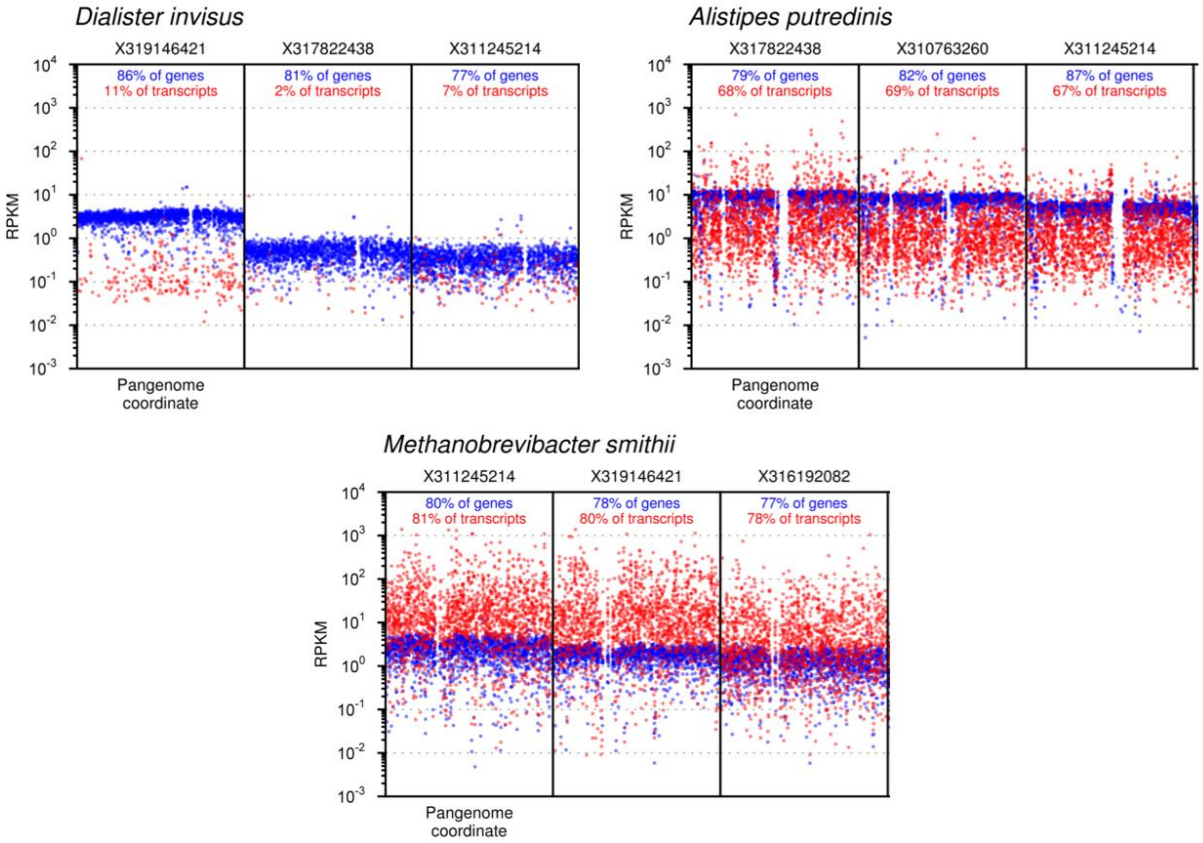


FIGURE S14. Comparison of species-specific transcription for two microbial growth-related pathways. Ribosomal protein transcription and peptidoglycan biosynthesis transcription were strongly correlated across species (Spearman's *r*=0.78; two-tailed *P*<<0.001). While the majority (81%) of species were over-transcribing ribosomal proteins, the majority (76%) of species were under-transcribing the peptidoglycan biosynthesis pathway. This is consistent with the community-level pattern observed in Fig. 4 of the main text. Within species, ribosomal protein biosynthesis tended to be more highly transcribed than the peptidoglycan biosynthesis pathway, possibly due to the different biological roles of the end-products of these two pathways (physical building blocks versus enzymes, respectively). This analysis was based on the pan-genome mapping procedure described in the Supporting Methods section. Within each species, the abundance values of genes implicated in the ribosomal protein and peptidoglycan biosynthesis pathways were summed at the RNA and DNA levels. The log ratio of these sums for a particular pathway provided an approximate measure of the pathway's per-species transcriptional activity (relative to its metagenomic abundance).

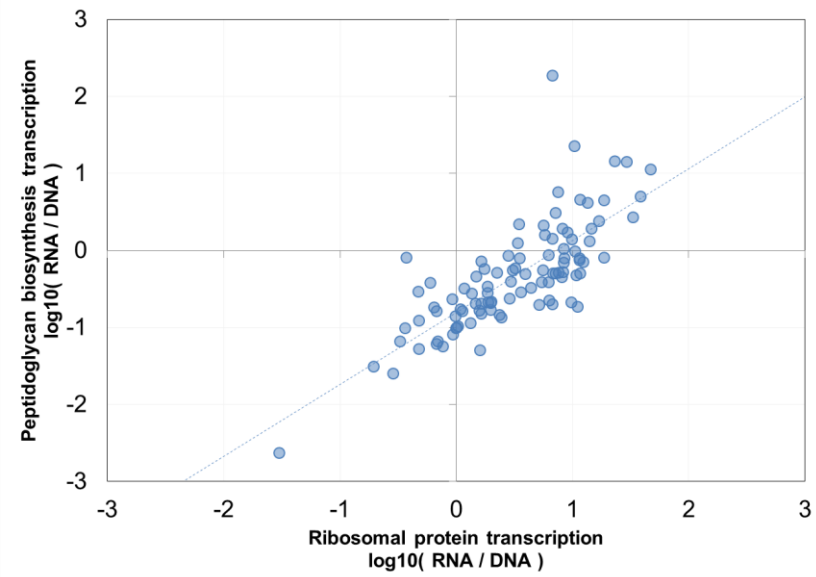


FIGURE S15. Between-subject, between-method analysis of variably transcribed pathways. For the 4 most DNA-abundant genes from each of the variably transcribed pathways highlighted in Fig. 6, we log-transformed the genes' relative RNA abundances and then converted these values to *z*-scores calculated over the 8 subjects. This was repeated for each sample handling method. In addition to highlighting between-subject variation in pathway expression, this figure highlights (i) consistency of pathway member-gene expression within subjects (e.g. expression levels for the 4 ribosomal genes tend to rise and fall together) and (ii) consistency of within-subject pathway variation across sample handling methods (e.g. chaperones are up-regulated in Subject 1 across the 3 sample handling methods). The fourth example of a proteasome-associated gene from Fig. 6 (proteasome accessory factor A) was rarely detected at the RNA level and was excluded from this analysis; NAs indicate additional values below the limit of detection.

z-score(log10(RNA abundance)) FROZEN Samples								z-score(log10(RNA abundance)) ETOH-Fixed Samples								z-score(log10(RNA abundance)) RNAlater-Fixed Samples								Functional module	Gene
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8		
-1.8	-0.3	-0.4	-0.6	1.3	1.1	0.5	0.2	-1.1	-0.7	1.2	-1.0	0.7	1.3	0.3	-0.8	-1.7	0.5	1.4	-0.4	0.0	1.1	-0.7	-0.2	Bacterial ribosome	K02886: large subunit ribosomal protein L2
-1.1	-0.2	-1.4	-0.4	0.4	1.6	0.9	0.2	-1.0	-1.3	0.5	-0.5	0.5	1.7	0.6	-0.6	-1.2	-0.2	0.5	-0.3	0.2	2.1	-0.8	-0.3	Bacterial ribosome	K02992: small subunit ribosomal protein S7
-1.7	-0.8	-0.2	-0.4	1.2	1.3	0.6	0.0	-0.7	-1.3	0.3	-0.6	0.9	1.7	0.5	-0.7	-1.6	0.1	1.7	-0.4	0.6	0.6	-0.8	-0.2	Bacterial ribosome	K02948: small subunit ribosomal protein S11
-1.6	-0.8	0.8	-0.7	1.2	1.0	-0.2	0.3	-1.7	-0.4	0.9	-1.1	0.4	1.2	0.3	0.4	-1.6	-0.3	1.3	-0.4	0.5	1.1	-1.0	0.4	Bacterial ribosome	K02931: large subunit ribosomal protein L5
1.7	-0.9	-0.2	0.0	1.4	-0.7	-0.4	-0.9	2.3	-0.7	0.2	0.1	-0.5	-0.1	-0.2	-0.9	2.2	0.0	-0.4	0.6	-1.1	-0.6	-0.1	-0.6	Chaperones	K03686: molecular chaperone DnaJ
1.5	-0.8	-0.3	0.2	1.5	-0.7	-0.3	-1.1	1.9	-0.9	0.8	-0.2	0.0	-0.4	0.1	-1.3	2.1	-0.1	-0.1	0.5	-0.6	-1.0	0.1	-0.8	Chaperones	K04077: chaperonin GroEL
1.8	-0.8	-0.3	0.2	1.2	-0.7	-0.6	-0.9	2.2	-0.2	0.5	-0.3	-0.3	-0.6	-0.1	-1.2	2.2	0.0	-0.2	0.6	-0.9	-0.9	-0.2	-0.6	Chaperones	K04043: molecular chaperone DnaK
1.6	-0.7	0.6	-0.3	1.2	-0.9	-0.5	-1.0	1.5	-0.9	1.0	0.2	0.0	-0.8	0.4	-1.4	2.0	-0.1	-0.3	0.8	-0.2	-1.1	0.0	-1.0	Chaperones	K04078: chaperonin GroES
0.2	0.2	-0.5	-0.3	0.7	1.9	-0.8	-1.4	0.1	-0.8	1.3	-0.5	0.4	1.1	0.2	-1.7	0.3	0.6	-0.2	-0.7	1.4	0.9	-0.7	-1.6	Uronic acid metabolism	K01686: mannonate dehydratase [EC:4.2.1.8]
-0.3	0.3	-1.3	0.2	0.9	1.6	-0.3	-1.2	-0.3	-0.2	0.6	-0.5	0.9	1.6	-0.3	-1.7	-0.6	0.1	-0.5	0.0	0.9	1.9	-0.5	-1.3	Uronic acid metabolism	K01685: altronate hydrolase [EC:4.2.1.7]
0.9	0.0	-0.9	-1.6	0.9	1.3	-0.6	-0.1	0.5	-1.0	0.7	-1.2	1.2	0.9	0.0	-1.2	0.8	-0.4	0.4	-1.5	0.8	1.4	-1.2	-0.2	Uronic acid metabolism	K01812: glucuronate isomerase [EC:5.3.1.12]
-0.3	-0.1	-0.6	-1.0	1.8	1.3	-0.4	-0.8	-0.3	-0.8	1.0	-1.0	1.3	1.1	0.0	-1.1	-0.6	-0.3	0.4	-1.0	1.4	1.5	-0.8	-0.8	Uronic acid metabolism	K00041: tagaturonate reductase [EC:1.1.1.58]
-1.3	0.8	0.7	0.2	0.0	-1.4	NA	1.1	-0.9	0.9	0.5	0.8	0.3	-1.5	-1.1	1.0	-1.0	0.9	0.1	0.6	0.5	NA	-1.8	0.6	Bacterial proteasome	K03432: proteasome alpha subunit [EC:3.4.25.1]
-1.5	0.9	0.7	0.4	0.1	-0.8	-1.0	1.2	-1.5	1.0	0.0	0.5	0.2	-1.2	NA	1.0	-1.8	0.6	0.1	0.7	0.7	-1.0	NA	0.7	Bacterial proteasome	K03433: proteasome beta subunit [EC:3.4.25.1]
-1.2	0.4	1.1	-1.6	0.7	0.3	NA	0.1	-1.8	0.1	1.4	0.3	-0.7	0.3	NA	0.3	-0.5	0.4	1.1	1.0	-1.3	0.7	-1.4	-0.1	Bacterial proteasome	K13527: proteasome-associated ATPase

Scale	-2.0	-1.0	0.0	1.0	2.0

Scale	-2.0	-1.0	0.0	1.0	2.0

Scale	-2.0	-1.0	0.0	1.0	2.0

Legends for Supporting Datasets

Dataset S1. (Excel spreadsheet) Results of two-way ANOVA evaluating subject and sample handling method effects on microbial transcripts. Less than 5% of total transcripts experience a strong, statistically significance sample handling effect. See main text for description of statistical analysis procedures.

Dataset S2. (Excel spreadsheet) Over- and under-expression of genes in the gut microbiome. Log RNA/DNA relative abundance ratios were computed for each subject. The mean of these values across subjects was tested for significant deviation from 0 using a one-sample *t*-test.

Dataset S3. (Excel spreadsheet) Pathways enriched for over- or under-expressed genes. KEGG pathways and modules were treated as unstructured gene lists and compared to the ranked list of genes in Dataset S2 to identify significant enrichments.

Dataset S4. (Excel spreadsheet) Genes ranked by relative expression variability. We computed the coefficient of variation for each gene at the RNA level and the DNA level. The ratio of these two values provides a measure of the relative degree of between-subject variation in the gene's expression.

Dataset S5. (Excel spreadsheet) Pathways enriched for variably-expressed genes. KEGG pathways and modules were treated as unstructured gene lists and compared to the ranked list of genes in Dataset S4 to identify significant enrichments.

Dataset S6. (Binary Excel spreadsheet) Results of the pan-genome mapping analysis. Columns are labeled as "SUBJECTID_Whole_DNA" for the 8 control stool metagenomes and "SUBJECTID_Whole_RNA" for the 8 control stool metatranscriptomes. Each row corresponds to one gene, and is annotated with species name, IMG gene ID, and colon-delimited KEGG Orthogroup assignments (if any exist). Values reported in the table are count data representing read hits to these genes. See Supporting Methods for details.

References

1. Segata N*, et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9(8):811-814.
2. Abubucker S*, et al.* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 8(6):e1002358.
3. Markowitz VM*, et al.* (2012) IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* 40(Database issue):D123-129.
4. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.
5. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.
6. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.
7. The Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature* 486(7402):215-221.
8. Giannoukos G*, et al.* (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* 13(3):R23.