Title

Metatranscriptome of human fecal microbial communities in a cohort of adult men

Supplementary information – SI

Galeb S. Abu-Ali[a,c,1], Raaj S. Mehta[d,1], Jason Lloyd-Price[a,c], Himel Mallick[a,c], Tobyn Branck[a,g], Kerry L. Ivey[b,h], David A. Drew[d,e], Casey DuLong[a], Eric Rimm[b,i], Jacques Izard[f], Andrew T. Chan[c,d,e,i,1,2], Curtis Huttenhower[a,c,1,2]

[a]Biostatistics Department and [b]Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [c]The Broad Institute, Cambridge, MA 02142; [d]Clinical and Translational Epidemiology Unit, [e]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114; [f]University of Nebraska, Lincoln, 1901 North 21 Street, Lincoln, NE 68588; [g]U.S. Army Natick Soldier Systems Center in Natick, MA 01760, [h]South Australian Health and Medical Research Institute, Infection and Immunity Theme, School of Medicine, Flinders University, Adelaide, Australia, 5000; [i]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital

[1]Contributed equally to this work

[2]To whom correspondence should be addressed:

Curtis Huttenhower
Harvard T. H. Chan School of Public Health
Department of Biostatistics
655 Huntington Avenue
Building 1, Room 413
Boston, Massachusetts 02115
Phone: 617.432.4912
chuttenh@hsph.harvard.edu

Andrew T. Chan
Massachusetts General Hospital
Clinical and Translational Epidemiology Unit
55 Fruit St, GRJ-825C
Boston MA 02114
achan@mgh.harvard.edu

Supplementary Notes

Viruses detected from non-enriched metagenomes and metatranscriptomes do not reflect variation of bacterial taxa

A small number of both DNA and RNA viruses were quantified confidently by MetaPhlAn2, which is likely an underestimate of the gut virome diversity since our extraction protocol did not enrich for virus-like particles. Of the 30 DNA viral species detected in the cohort, 29 were bacteriophage belonging primarily to the Siphoviridae (19) family, with few Myoviridae (6), Podoviridae (2), and Inoviridae (1) members, and one *Escherichia* phage of unknown taxonomy. The most common putative phage identifications were *C2likevirus* (in 116 participants), *Epsilon15likevirus* (in 16 participants), and *Lactobacillus Lc Nu* (in 16 participants), however, no correlation was found between the abundance distribution of these Caudovirales phages and their natural host genera *Lactococcus*, *Escherichia*, and *Lactobacillus*; although power for detecting this is low given non-virally-enriched detection rates. In metatranscriptomes, we identified 88 RNA viral species belonging to 19 families. Apart from rare Leviviridae and Iflaviridae members, these were all plant viruses, in agreement with previous studies[1] and possibly due to dietary ingestion. Most RNA viruses were detected in <30 (10%) metatranscriptomes with the exception of *Pepper mild mottle virus* (48% prevalence) and *Tomato mosaic virus* (39% prevalence), which together accounted for 31% of viral RNA on average (when present). Although gut viral ecology is more difficult to analyze than that of the bacteriome due to inadequate viral reference sequences[2], these methods allow for some incidental analysis of DNA phage and RNA plant viruses in human gut metagenomes and metatranscriptomes.

Effect of GC content and ORF length on transcription ratios.

We analytically evaluated the effect of GC content and ORF length on transcription ratios, finding no interaction (**Supplementary Fig. 5**). The 430 structured MetaCyc pathways analyzed here were quantified from 808,694 UniRef90 gene families (3.4% of the total UniRef90 database) that had detectable DNA in at least one sample in our study. Of those, 89,991 and 44,792 UniRef90s had non-zero DNA and RNA abundance in at least one sample, respectively. This resulted in 37,085 pathway-associated UniRef90 gene families for which RNA/DNA ratios were calculated. Among these genes, there were no significant differences in ratio when stratified by either %GC or length of nucleotide sequences; when tested continuously, there was no significant correlation of either length or %GC with RNA abundances. Intriguingly, a very low effect size (Pearson's r<-0.04), but significant, correlation was observed between sequence length and DNA abundance. Given the comparability of DNA and RNA protocols in this study, it is not clear why this might have arisen, but at such a minimal effect size it does not affect the study's conclusions.

Core and variable fecal metatranscriptomes differ from the metagenome.

The distribution of transcript abundance ranged over four orders of magnitude among 210 pathways that were transcribed in >10% of samples (**Supplementary Fig. 6F**). The highest transcription ratios consistently arose from pathways that were both low prevalence and taxonomically restricted, e.g. archaeal methanogenesis and coenzyme 420 biosynthesis, as previously suggested by our pilot study[3]. Following energy metabolism and fermentation, which tended to dominate in both prevalence and expression levels, the highest metatranscription was observed for saturated and unsaturated fatty acid elongation pathways, albeit in less than one-third of samples. Fatty acids are generated from acetyl-CoA, which in turn is produced mainly during glycolytic energy release, and together this may explain the concerted metatranscription of glycolysis and energy-expensive fatty acid elongation. As the primary role for bacterial fatty acids is to serve as precursors for cell membrane building blocks (e.g. phospholipids), this likely

signals widespread cell growth in the typical fecal microbiota[4]. On the other end of the spectrum, pathways with the lowest metatranscription had mean RNA abundances below their corresponding DNA relative abundances, with prevalence of metatranscription ranging between 15% (sulfate assimilation/cysteine synthesis) and 95% (peptidoglycan synthesis) of samples. This low tail of metatranscription included several amino acid synthesis pathways, including methionine, homoserine, aromatic and seleno-amino acids, followed by cofactor biosynthesis, including thiamine (and variants), tetrapyrrole, etc. Prevalent metatranscription of degradation of stachyose (PWY-6527), a legume-derived non-digestible tetrasaccharide that promotes SCFA producers, may reflect diet preferences. Together, these findings would underline that the fecal microbiome does not prioritize *de novo* synthesis of amino acids or widespread activation of specialized functions, yet displays high dynamic range and milieu activities such as transformation of phenolics, stress adaptation, and secondary metabolism.

Genetic divergence patterns of stool-associated bacterial strains is species-specific and preserved among host populations

Nucleotide substitution rates within and between cohorts were strikingly similar for the compared species, indicating that species' evolutionary strategies within the stool niche were comparable between these host populations (**Fig. 6C**). The amount of genetic change was higher for Firmicutes than Bacteroidetes and did not appear to be simply a function of species prevalence in the two cohorts. For example, *Bacteroides dorei* and *uniformis*, and *Alistipes putredinis* had comparable prevalence with *Ruminococcus bromii*, *Dialister invisus*, and *Eubacterium rectale*, yet appreciably fewer nucleotide substitutions between strains. This may be due to *Bacteroides* species' more restrictive definitions by systematics[5], serving as a reminder that culture-based isolate information and culture-independent microbial profiling may need further resolution as strain and transcriptional meta'omics are explored.

Species-function relationships in fecal meta'omes.

We quantified how tightly was each pathway coexpressed - that is, the extent to which the multiple enzymes making up each pathway were expressed at similar abundances within each organism and meta'ome (**Supplementary Dataset 4**). This was assessed using the average variance of gene families' transcription log ratios across samples (see **Methods**), here termed the EC dispersion. The distribution of dispersions from all pathways' ECs was significantly below 1 (one-sided t-test $P=1.1\times10^{-16}$), with a mean of 0.89, indicating that functionally-related genes are co-expressed on average. Tightly coexpressed pathways (low dispersion) included methanogenesis (dispersion 0.26), two pathways for L-histidine degradation (0.38, 0.39), and degradation of the glutaryl-CoA (0.49) intermediate of tryptophan and lysine metabolism. Tryptophan and histidine are among the energetically most expensive amino acids to synthesize[6,7], for which tight co-expression of degradation pathway is not surprising. No evidence was found for a relationship between EC dispersion and the number of species that transcribed the pathway (Spearman rho -0.01). Differences between pathways that were considered a part of the core or variably expressed metatranscriptome were also not detected (Wilcoxon rank-sum test p-value 0.10).

Supplementary Discussion

We briefly review here the current literature on the topic of microbiome sample stabilization with RNAlater. The reported minor effect of choice of sample handling method on microbiome composition[8] lacks testing for statistical significance of any variance, suggesting that between-condition variation in that study was comparable to replicate variation (and much smaller than population variability). In addition to our own validation work[3], which indicates a negligible effect of RNAlater on microbial community composition, there are numerous reports on the evaluation of methods for storage and handling of microbiome samples in a cohort setting. These studies reveal a lack of significant alteration in community structure between samples preserved with RNAlater, ethanol, lyophilization, fecal occult blood test cards, and freezing at -80℃ [9-14].

The recent Choo et al. study[15] reports a statistically significant effect of storage method on microbial community composition, based on assessing the variation among differently stored samples collected from one individual sampled three times over 30 days. However, Choo et al. report that the variation attributable to storage method is markedly smaller than the variation explained by different sample time points (i.e. smaller than intra-individual difference, which in turn is far smaller than inter-individual difference). Notably, the variation introduced by RNAlater, relative to freezing at -80℃, was comparable to that introduced by sample storage in OMNIGene.Gut, another popular sample stabilization kit. Furthermore, significant but localized differences in taxon abundance relative to freezing were comparable among RNAlater and OmniGene.Gut. Finally, Choo et al evaluated stool samples from a single subject, which are not likely to be representative at the population level.

Supplementary Figures

Please note that **Supplementary Figures 6-10** are multi-panel, multi-page figures that were submitted as separate files due to their length and size.

Supplementary Figure 1. Taxonomic profiles of gut metagenome ecology and stability. Summaries of taxonomic membership and population diversity in the MLVS, which broadly agree with previous comparable gut metagenome profiles[16-18]. **A)** Inter-individual variation of major phyla. Seven out of fourteen phyla were found present at a relative abundance >0.1% with a prevalence >10%. Panels indicate collection time points, with the number of participants in parenthesis, and samples ordered by decreasing mean abundance of the most abundant phylum (Firmicutes). **B)** Relative abundances of most abundant (when present) species (rows) across 913 samples. Columns in the heatmap were ordered based on average linkage clustering on a Euclidean distance matrix of $\log_{10}$ relative abundances. The grey color indicates that the species was not detected. **C)** Principal coordinates ordination of 307 subjects on Bray-Curtis dissimilarity between abundance profiles for 139 species (detected in ≥10% samples at ≥0.01% abundance) averaged over time points and colored by sequencing depth, and each point in the ordination is one participant. The ordination of taxonomic profiles averaged for individual time points shows no bias from variability in input sequencing depth of samples. Labels t1-t4 represent sampling time points; samples were self-collected in two pairs (t1-t2 and t3-t4), six months apart, with each pair spanning 2-3 days.

Supplementary Figure 2. Variation in genus composition among MLVS, HMP1-II, and ELDERMET cohorts. PCoA of Bray-Curtis dissimilarities in genus composition of samples from (**A**) the ELDERMET cohort (participant age >65 yrs) reported by Claesson *et al*[19], (**B**) the ELDERMET, MLVS (age range 65-81 yrs) and HMP1-II (age range 20-40 yrs) cohorts based on the intersection of genera detected in all three cohorts, and (**C**) the ELDERMET, MLVS and HMP1-II cohorts based on the union of *all* genera detected in the three cohorts. Cohort sample numbers are in parenthesis. Only genera with abundances >=0.01% in >=10% of samples in respective cohorts were used for ordination analysis; i.e. 42, 57, and 52 genera in ELDERMET, MLVS, and HMP1-II cohorts. **A)** Taxonomic composition of the ELDERMET cohort reported in [19] was recapitulated in our analysis indicating that 72% samples were dominated by Bacteroidetes with 56% average abundance across all 192 samples; Firmicutes averaged 39% abundance per sample. In contrast to Claesson *et al*[19], though, we found that the control samples (9 young adults) in the ELDERMET cohort also contained Bacteroidetes at a slightly higher average abundance (45%) than Firmicutes (43%). Such differences in resulting taxonomic profiles are likely a consequence of different OTU calling pipelines. Claesson *et al* used the RDP tool suite version 10.16 whereas we used UPARSE version 9.0.2132 for *de novo* OTU clustering and the RDP classifier version 2.2 for taxonomic classification of OTU centroid sequences against the Greengenes 13_8 database. **B)** Twenty-seven genera represented the taxonomic intersection of all three cohorts. In our MLVS data, taxonomic profiling reinforced Firmicutes and Bacteroidetes as the dominant provenance of bacterial clades, and, unlike in young adults of the HMP1-II cohort[20,17], the proportion was tilted in favor of Firmicutes (50.6% ± 14.3%; mean ± s.d.) over Bacteroidetes (40.4% ± 14.3%) in the MLVS. One potential confounder in this comparison is the different DNA extraction protocol between MLVS and HMP, which was more efficient in extracting Bacteroidetes DNA from HMP samples[21,22]. **C)** Seventy genera represented the taxonomic union of all three datasets and, in addition to sample processing, the comparison is biased by differences in taxonomic assignment strategies for metagenomic and amplicon sequencing reads. Explicit comparison with data from an ELDERMET publication including whole metagenome shotgun (WMS) sequence data[18] was not possible as neither WMS nor 16S sequencing reads from that study are available from the MG-RAST server where the data were deposited

(http://metagenomics.anl.gov/mgmain.html?mgpage=project&project=mgp154). Taken together, these larger metagenomic and new metatranscriptomic data showed a greater enrichment for Firmicutes, clearer resolution at the species level, and fewer signs of instability or directly age-linked configurations, possibly due to the comparably high level of population health despite participant age. All numbers in parenthesis are stool metagenome sample counts from a total of 307 MLVS, 253 HMP (male and female), and 170 ELDERMET (male and female) participants.

Supplementary Figure 3. Feature detection as a function of sequencing depth. Effect of sample sequencing depth on the ability to detect microbiome functional features in metagenomic and metatranscriptomic sequence data. HUMAnN2 functional profiling of DNA and RNA quality filtered reads was performed on individual samples in species-specific mode, i.e. nucleotide alignment against pangenomes of species identified in the sample with MetaPhlAn2, and in combined species-specific and -agnostic mode, in which reads not matching any pangenome reference sequences were subjected to DIAMOND[23] translated searching against the UniRef90 database. Each sample is represented by a green and blue point in each plot. Linear regression trends with 95% confidence intervals are represented by straight lines and grey shading in each plot. Four plots per row from left to right show read alignment rates, and counts of detected UniRef90 gene families, enzymes, and pathways as a function aligned read counts, for 913 DNA samples (**A**) and 347 RNA samples (**B**). The number of gene families detected in metagenomic samples increased by less than half a log over a log difference in sequencing depth (**A**), but well over one log for metatranscriptomes (**B**) indicating great transcriptional capacity of the gut metagenome. Detection of UniRef90 transcripts Increasing sequencing depth would improve feature detection from RNA samples whereas feature detection was saturated with these input DNA read counts. **C)** Species rarefaction curve for samples with total counts above the 1st decile (836 samples). The vertical reference line is set at 50,000; at rarified count of 50,000 the median ratio of rarified to observed number of species of samples nears one (boxplot inset). On average, 3.5% of input reads per sample were considered by MetaPhlAn2, implying per sample saturation at <1.5M input reads which is roughly seven times less than the average per sample sequencing depth after quality control (9.3M paired-end reads). The ratio of the Chao extrapolated richness from all samples ('specpool' function in the R/vegan package) to the observed number of species in all samples, indicated 89.9% species saturation in the MLVS cohort. **D)** Rarefaction curves of UniRef90 gene family abundances, using data from samples with total counts above the 9th decile (93 samples), plateaus at a count of 5M. This was also indicated by boxplot summaries of rarefied to observed UniRef90 ratios (inset). The per-sample average read usage rate by HUMAnN2 was 60%, implying per sample saturation at 8.3M quality filtered reads. **E)** For UniRef90 transcript abundances, curves plateau at similar rarefaction levels based on analysis of 184 samples with highest total counts. Average sequencing depth for RNA samples was 6.7M paired-end reads, after quality control. Boxplot whiskers represent 1.5 times the inter-quartile range from the first and third quartiles. RPKs – reads per kilobase.

Supplementary Figure 4. Definition of core metatranscriptome that is robust to sequencing depth. Number of pathways (from a total of 340) with prevalence exceeding the given threshold, calculated from 341 samples with RNA sequencing depth greater than 1M, 2M, 4M, and 8M reads. A change in slope is observed at 81 pathways, which is robust to changes in sequencing depth. These pathways were thus defined as "core".

Supplementary Figure 5. UniRef90 gene and DNA-normalized transcript abundance is not biased by GC content and ORF length. A) RNA/DNA ratios for gene families (UniRef90s, total n=37,085) do not vary significantly by GC content, plotted as deciles from the lowest to highest %GC in gene families analyzed in 341 metagenome-metatranscriptome paired samples from 96 MLVS participants. GC content was calculated as an average across a single representative

nucleotide sequence per UniRef90 family. Boxplot whiskers represent 1.5 times the inter-quartile range from the first and third quartiles. Half-open interval labels for x-axis ticks include only the second endpoint, e.g. (0,1] includes values greater than zero and less than or equal to 1. **B)** When analyzed continuously (rather than quantized), neither DNA abundance nor RNA abundance of each UniRef90 gene family (one per point) is strongly correlated with %GC. Additionally, this small degree of %GC bias does not differ between DNA vs. RNA abundances. **C)** As above, RNA/DNA ratios for gene families do not vary significantly by gene length, as deciles from highest to lowest lengths across gene families. Length was again calculated using a single representative sequence per UniRef90. Boxplot whiskers represent 1.5 times the inter-quartile range from the first and third quartiles. **D)** As above, neither DNA nor RNA abundance of UniRef90 gene families were strongly correlated with length. A slightly greater association was detected between greater metagenomic (but not metatranscriptomic) abundance of shorter genes, but even this weak association was of extremely low effect size (maximum absolute Pearson correlation <0.04).

Supplementary Figure 6. Core and variable metatranscriptomes of the stool microbiome, with pathway definitions and distribution range of pathway transcript abundances. DNA-normalized transcript abundances for 239 gut microbiome pathways with detectable RNA in >10 of the 341 metatranscriptomes, collected from 96 MLVS participants. **A)** Core metatranscriptome pathways (transcribed in >80% of samples) with RNA:DNA transcription ratio >1. **B)** Low-expression core metatranscriptome pathways with transcript abundance detectable in >80% of samples but an RNA:DNA ratio <1. **C)** Variably metatranscribed pathways detected in DNA but below detection in at least half of RNA samples, and **D)** variably metatranscribed pathways below detection in DNA (and matching RNA) in 30%-80% of the 341 samples. **E)**, Thirty-eight pathways that do not fall into any of the categories depicted in **A-D**. **F)** Pathways with the 30 highest and 30 lowest mean DNA-normalized transcript abundances among the 341 metatranscriptome samples. Points indicate individual samples with medians overlaid per pathway, with prevalence in parenthesis; see Supplementary Notes for supplementary results text.

Supplementary Figure 7. Per pathway species contributions to metagenomes and metatranscriptomes. Each point in a given pathway plot is a contributing species, and species contributions to DNA and RNA are expressed as relative abundances; i.e. the average abundances from 341 metagenome-metatranscriptome sample pairs from 96 participants. For example, if DNA or RNA for a given pathway is contributed by a single species, based on species-specific HUMAnN2 functional profiling, then the corresponding log10 value along the x or y axis, respectively, is 1. Color scheme: red - species (points) that contributed more RNA than DNA for a given pathway; blue - species (points) that contributed more DNA than RNA for a given pathway; grey - species (points) that contributed equal levels of RNA and DNA for a given pathway. Number of species within each plot (n) and Spearman correlation coefficients (Rho) between species' contributions to DNA and RNA abundances of a pathway are provided in plot titles.

Supplementary Figure 8. Species-stratified distributions of metagenomic potential (DNA) and metatranscriptomic activity (RNA) for all pathways with non-zero abundance in at least 10% of samples. The 40 most transcriptionally-active species are shown (additional species are grouped as "other"). Abundances were normalized within each pathway for 189 subject-week pairs, from 96 participants. For each pathway, the number of samples with non-zero RNA and DNA is given in the x-axis label. Subjects were ordered to emphasize blocks of subjects with similar metatranscriptomic profiles (see **Methods**). Pathways are sorted in decreasing order of their Weighted Spearman coefficients (see **Fig. 4B**).

Supplementary Figure 9. Ecological interactions in the gut microbiome for individual time points. Significant co-variation and co-exclusion relationships among 104 species in stool metagenomes of MLVS participants. Each node represents a species and edges correspond to

significant interactions inferred by BAnOCC (see **Methods**). Stool microbiome taxonomic profiles were averaged within each participant for the first (215 participants) and second (258 participants) collection pairs (separated by 6 months). 95% credible interval criteria was used to assess significance, and only estimated absolute correlations with effect sizes >=0.15 are reported.

Supplementary Figure 10: Strain-level diversity is robust across cohorts. Principal coordinate analysis of pairwise nucleotide substitution rates among strains of 21 species identified in stool metagenomes from MLVS and HMP1-II cohorts. Nucleotide substitution rates were calculated from multiple sequence alignments using the Kimura Two-Parameter distance[24]. All numbers in plot titles are sample counts in which indicated strains were above limit of detection; from a total of 913 MLVS stool metagenomes and 564 HMP stool metagenomes (from 253 male and female HMP participants) that were analyzed with StrainPhlAn.

Supplementary Tables

Supplementary Table 1. Functional profiling of MLVS metagenomes and metatranscriptomes. UniRef90 gene families identified from DNA and RNA, plus those characterizable to enzymes and pathways per sample and in the entire cohort.

	Metagenome (*n* = 913)		**Metatranscriptome (*n* = 347)**	
	Features per sample (mean ± s.d.)	Unique in cohort*	Features per sample (mean ± s.d.)	Unique in cohort
UniRef90	173,609 ± 36,157	1,569,171	32,279 ± 21,537	602,896
ECs	1045 ± 128	1,909	623 ± 149	1,570
Pathways	253 ± 40	429	129 ± 48	340

* - Number of unique non-redundant features in entire cohort.

SI References

1 Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**, e3, doi:10.1371/journal.pbio.0040003 (2006).

2 Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157**, 142-150, doi:10.1016/j.cell.2014.02.032 (2014).

3 Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2329-2338, doi:10.1073/pnas.1319284111 (2014).

4 Zhang, Y. M. & Rock, C. O. Membrane lipid homeostasis in bacteria. *Nature reviews. Microbiology* **6**, 222-233, doi:10.1038/nrmicro1839 (2008).

5 SHAH, H. N. & COLLINS, M. D. Proposal To Restrict the Genus Bacteroides (Castellani and Chalmers) to Bacteroides fragilis and Closely Related Species. *International Journal of Systematic and Evolutionary Microbiology* **39**, 85-87, doi:doi:10.1099/00207713-39-1-85 (1989).

6 Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3695-3700, doi:10.1073/pnas.062526999 (2002).

7 Bender, R. A. Regulation of the histidine utilization (hut) system in bacteria. *Microbiology and molecular biology reviews : MMBR* **76**, 565-584, doi:10.1128/MMBR.00014-12 (2012).

8 Reck, M. *et al.* Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics* **16**, 494, doi:10.1186/s12864-015-1694-y (2015).

9 Song, S. J. *et al.* Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* **1**, doi:10.1128/mSystems.00021-16 (2016).

10 Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome biology* **16**, 73, doi:10.1186/s13059-015-0639-8 (2015).

11 Vlckova, K., Mrazek, J., Kopecny, J. & Petrzelkova, K. J. Evaluation of different storage methods to characterize the fecal bacterial communities of captive western lowland gorillas (Gorilla gorilla gorilla). *J Microbiol Methods* **91**, 45-51, doi:10.1016/j.mimet.2012.07.015 (2012).

12 Blekhman, R. *et al.* Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. *Sci Rep* **6**, 31519, doi:10.1038/srep31519 (2016).

13 Flores, R. *et al.* Collection media and delayed freezing effects on microbial composition of human stool. *Microbiome* **3**, 33, doi:10.1186/s40168-015-0092-7 (2015).

14 Vogtmann, E. *et al.* Comparison of Fecal Collection Methods for Microbiota Studies in Bangladesh. *Applied and environmental microbiology* **83**, doi:10.1128/AEM.00361-17 (2017).

15 Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci Rep* **5**, 16350, doi:10.1038/srep16350 (2015).

16 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

17 Consortium, H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).

18 Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178-184, doi:10.1038/nature11319 (2012).

19 Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proceedings of the National Academy of Sciences of the*

United States of America **108 Suppl 1**, 4586-4591, doi:10.1073/pnas.1000097107 (2011).

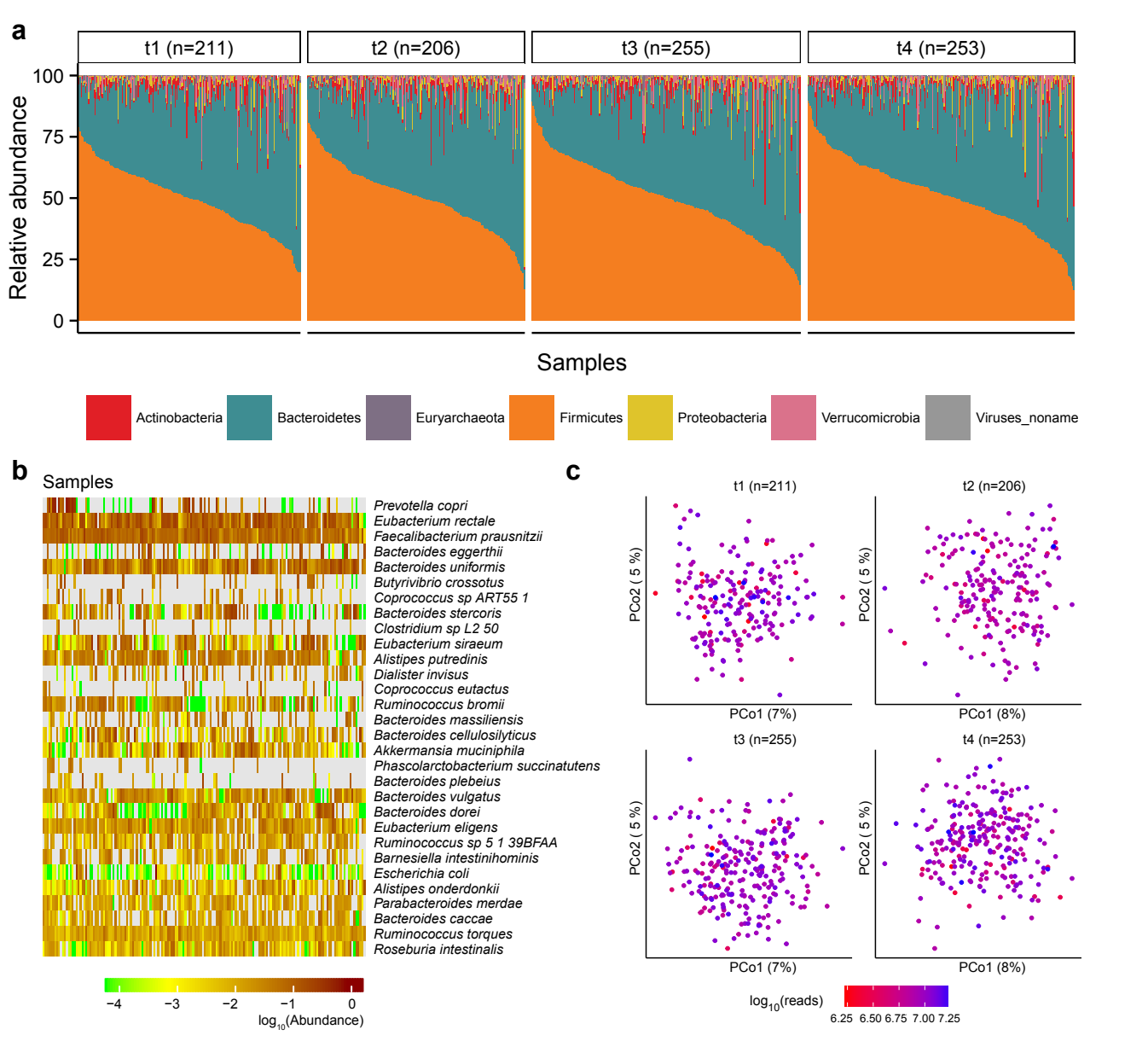
20 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

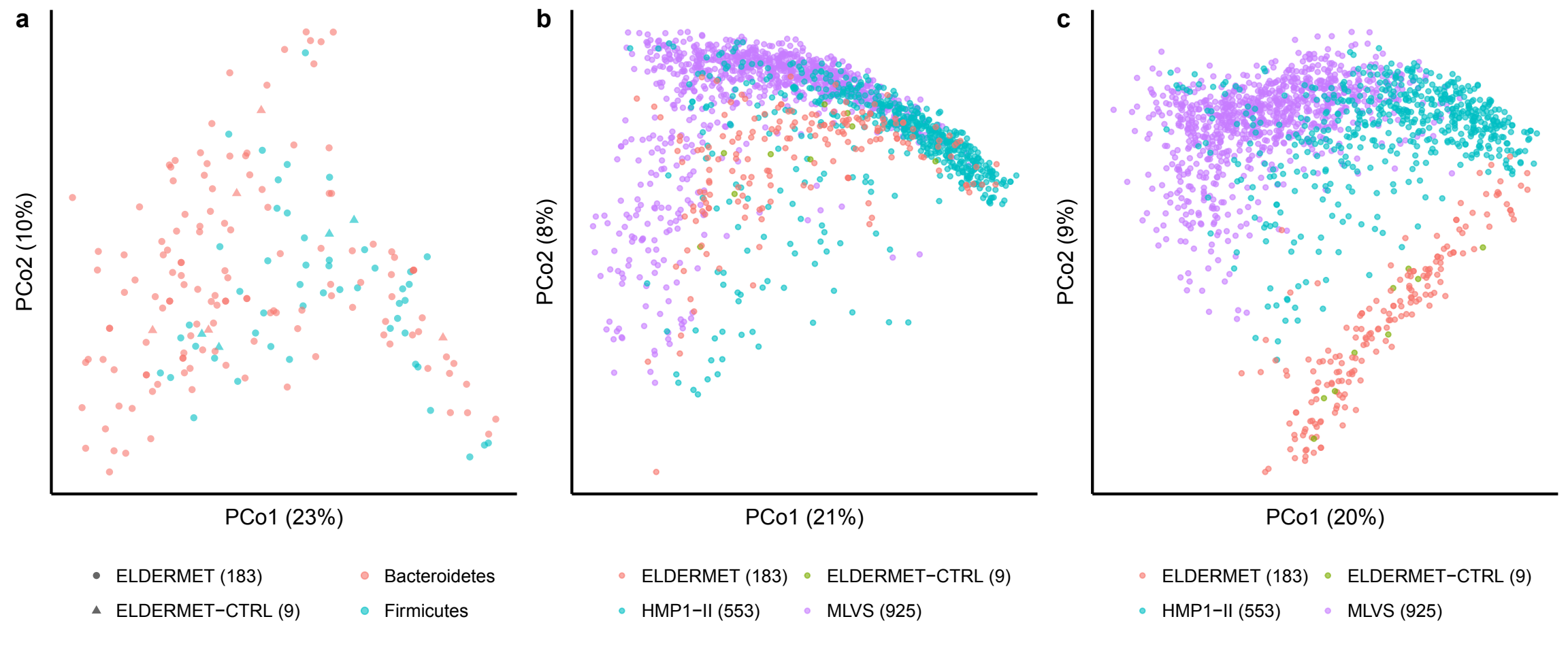
21 Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19, doi:10.1186/2049-2618-2-19 (2014).

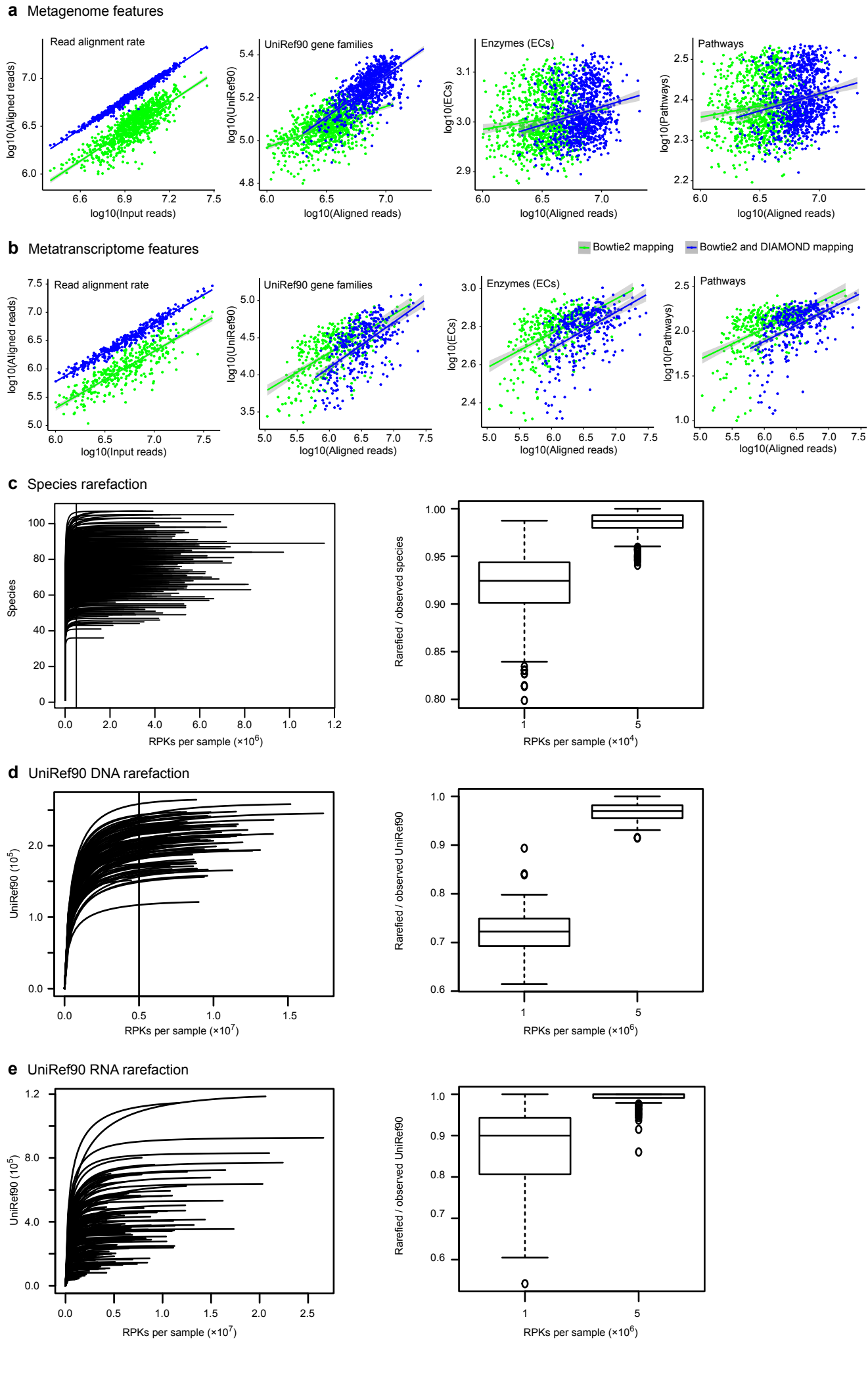
22 Shafquat, A., Joice, R., Simmons, S. L. & Huttenhower, C. Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends in microbiology* **22**, 261-266, doi:10.1016/j.tim.2014.01.011 (2014).

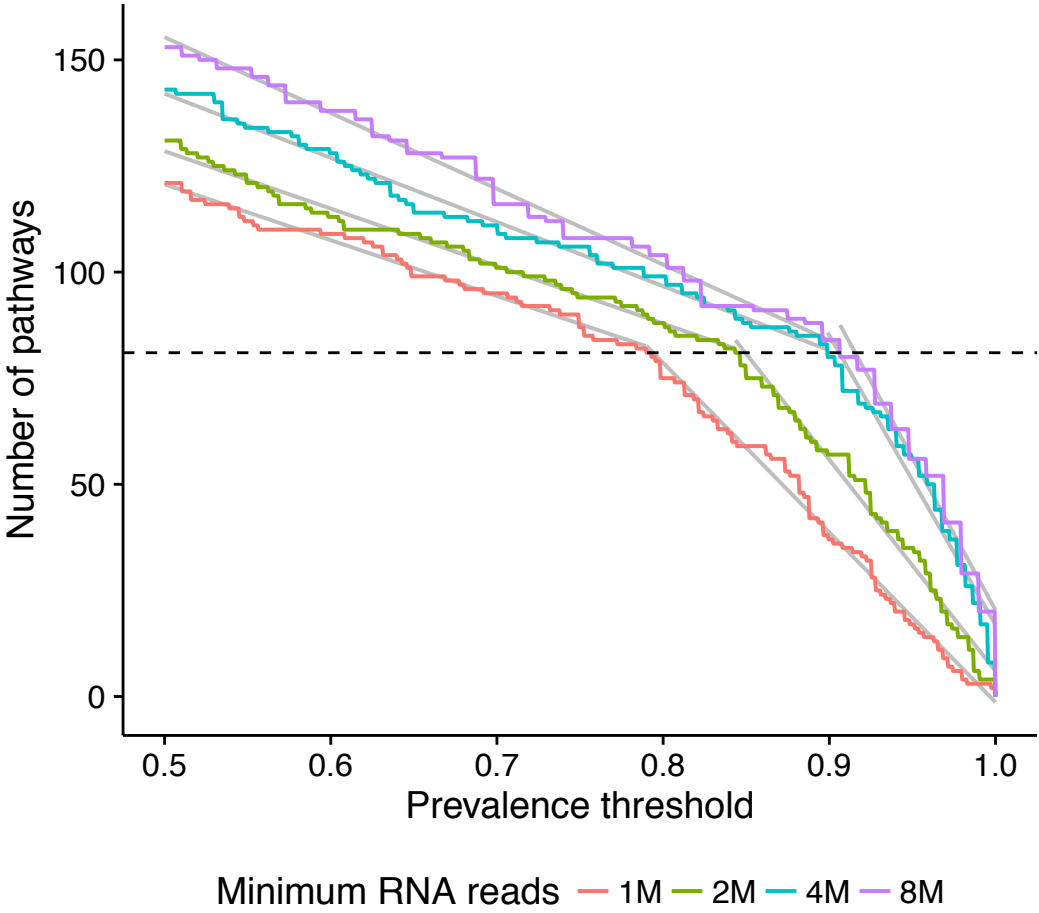
23 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

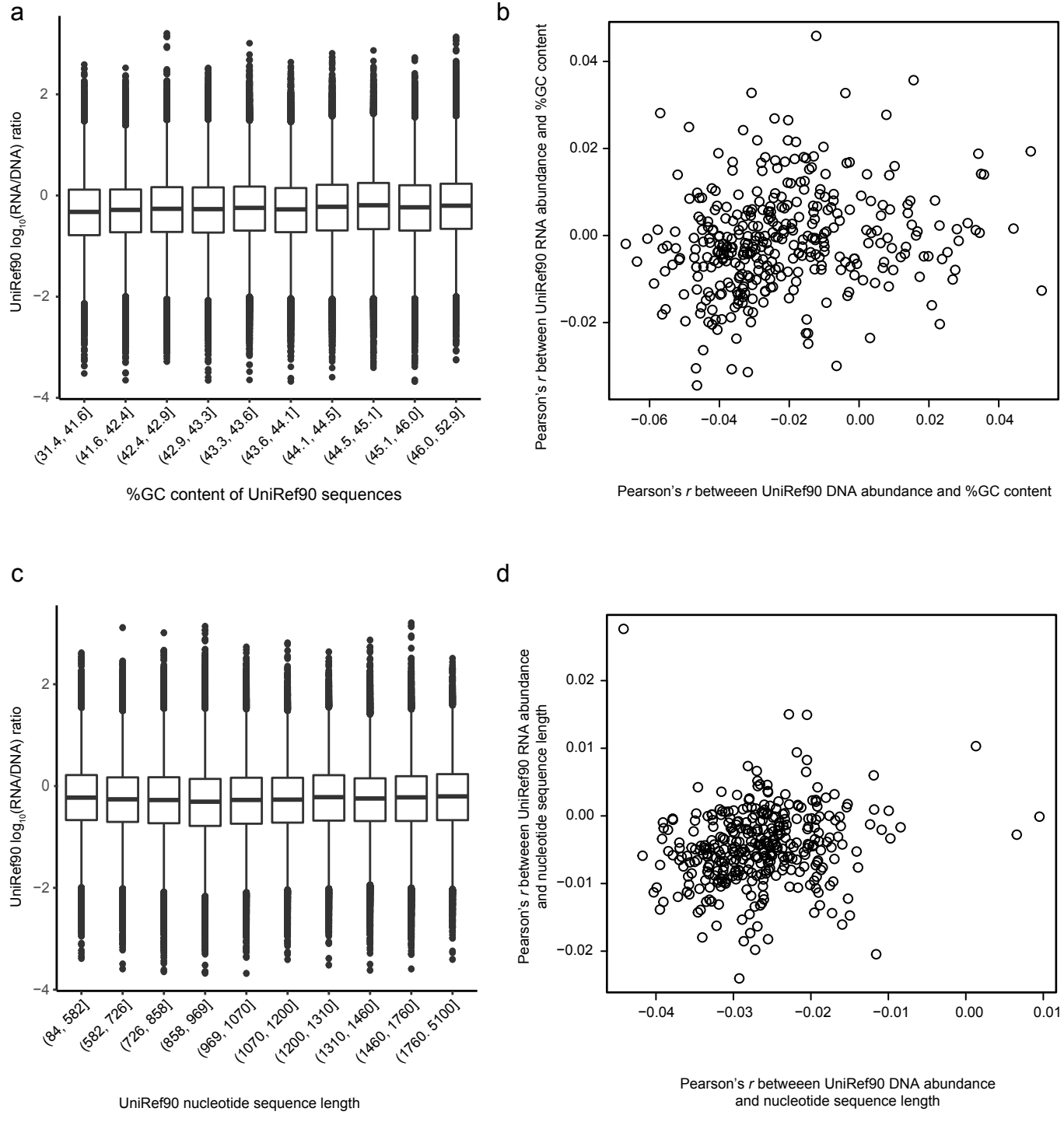
24 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980).


a
t1 (n=211)
t2 (n=206)
t3 (n=255)
t4 (n=253)
Relative abundance
100
75
50
25
0
Samples
Actinobacteria
Bacteroidetes
Euryarchaeota
Firmicutes
Proteobacteria
Verrucomicrobia
Viruses_noname
b
Samples
Prevotella copri
Eubacterium rectale
Faecalibacterium prausnitzii
Bacteroides eggerthii
Bacteroides uniformis
Butyrivibrio crossotus
Coprococcus sp ART55 1
Bacteroides stercoris
Clostridium sp L2 50
Eubacterium siraeum
Alistipes putredinis
Dialister invisus
Coprococcus eutactus
Ruminococcus bromii
Bacteroides massiliensis
Bacteroides cellulosilyticus
Akkermansia muciniphila
Phascolarctobacterium succinatutens
Bacteroides plebeius
Bacteroides vulgatus
Bacteroides dorei
Eubacterium eligens
Ruminococcus sp 5 1 39BFAA
Barnesiella intestinihominis
Escherichia coli
Alistipes onderdonkii
Parabacteroides merdae
Bacteroides caccae
Ruminococcus torques
Roseburia intestinalis
−4
−3
−2
−1
0
log10(Abundance)
c
t1 (n=211)
t2 (n=206)
t3 (n=255)
t4 (n=253)
PCo2 (5 %)
PCo1 (7%)
PCo1 (8%)
log10(reads)
6.25
6.50
6.75
7.00
7.25


a
PCo2 (10%)
PCo1 (23%)
ELDERMET (183)
ELDERMET−CTRL (9)
Bacteroidetes
Firmicutes
b
PCo2 (8%)
PCo1 (21%)
ELDERMET (183)
ELDERMET−CTRL (9)
HMP1−II (553)
MLVS (925)
c
PCo2 (9%)
PCo1 (20%)
ELDERMET (183)
ELDERMET−CTRL (9)
HMP1−II (553)
MLVS (925)


a Metagenome features
Read alignment rate
log10(Aligned reads)
log10(Input reads)
UniRef90 gene families
log10(UniRef90)
log10(Aligned reads)
Enzymes (ECs)
log10(ECs)
log10(Aligned reads)
Pathways
log10(Pathways)
log10(Aligned reads)
b Metatranscriptome features
Bowtie2 mapping
Bowtie2 and DIAMOND mapping
Read alignment rate
log10(Aligned reads)
log10(Input reads)
UniRef90 gene families
log10(UniRef90)
log10(Aligned reads)
Enzymes (ECs)
log10(ECs)
log10(Aligned reads)
Pathways
log10(Pathways)
log10(Aligned reads)
c Species rarefaction
Species
RPKs per sample (×10^6)
Rarefied / observed species
RPKs per sample (×10^4)
d UniRef90 DNA rarefaction
UniRef90 (10^5)
RPKs per sample (×10^7)
Rarefied / observed UniRef90
RPKs per sample (×10^6)
e UniRef90 RNA rarefaction
UniRef90 (10^5)
RPKs per sample (×10^7)
Rarefied / observed UniRef90
RPKs per sample (×10^6)


150
100
50
0
Number of pathways
0.5
0.6
0.7
0.8
0.9
1.0
Prevalence threshold
Minimum RNA reads
1M
2M
4M
8M


a
UniRef90 log10(RNA/DNA) ratio
2
0
−2
−4
(31.4, 41.6]
(41.6, 42.4]
(42.4, 42.9]
(42.9, 43.3]
(43.3, 43.6]
(43.6, 44.1]
(44.1, 44.5]
(44.5, 45.1]
(45.1, 46.0]
(46.0, 52.9]
%GC content of UniRef90 sequences
b
Pearson's r between UniRef90 RNA abundance and %GC content
0.04
0.02
0.00
−0.02
−0.06
−0.04
−0.02
0.00
0.02
0.04
Pearson's r betweeen UniRef90 DNA abundance and %GC content
c
UniRef90 log10(RNA/DNA) ratio
2
0
−2
−4
(84, 582]
(582, 726]
(726, 858]
(858, 969]
(969, 1070]
(1070, 1200]
(1200, 1310]
(1310, 1460]
(1460, 1760]
(1760, 5100]
UniRef90 nucleotide sequence length
d
Pearson's r betweeen UniRef90 RNA abundance and nucleotide sequence length
0.02
0.01
0.00
−0.01
−0.02
−0.04
−0.03
−0.02
−0.01
0.00
0.01
Pearson's r betweeen UniRef90 DNA abundance and nucleotide sequence length