Genomic variation landscape of the human gut microbiome: Individuality and temporal stability

Table of contents

1. Methods

1.1. Generation of a reference genome set

1,511 prokaryotic genomes were downloaded from GenBank and the MetaHIT Consortium (http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/) on 4 July 2010. Genes in these genomes were annotated with eggNOG[1] (version 2)

orthologous groups. A set of 40 universal single copy marker genes[2,3] was identified in these genomes using HMM profiles made for each marker gene from the corresponding orthologous group in eggNOG (hmmbuild and hmmsearch programs from HMMER2[4] were used; only the best hit was chosen since they were present in single copy in most genomes). For each marker gene, pairwise DNA sequence identities between all genomes were calculated using WU-BLAST[5] version 2.0 using parameters "`B=2000 spoutmax=3 span1`". For each genome pair, the median identity of all marker genes was used as a proxy for average nucleotide identity (ANI) between the two genomes. Using an operational 95% ANI recommended for identifying species[6], we generated 929 clusters of genomes (complete linkage). The clusters of genomes and selected reference genomes are listed in Supplementary Table 2.

1.2. Mapping of Illumina reads to reference genomes

Illumina reads from 266 (252 after quality control, see below) fecal metagenomes (124 from the European MetaHIT study[7], 139 from the US Human Microbiome Project[8], 3 obtained from Turnbaugh, P.J and Gordon, J.I., Washington University Center for Genome Sciences (NIH grant DK78669) part of a previously published study[9]; see Supplementary Table 1) were quality controlled using a customized trimming and filtering pipeline (described in ref. [10] section 5.2) with minor modifications. Briefly, bases were trimmed at the 5'-end of reads unless the number of base calls for any base (A, T, G, C) was within the average across all cycles plus/minus two standard deviations, bases were trimmed at the 3'-end reads if the quality score was <20, and reads shorter than 45 bp or reads with a median quality score < 20 were removed from further analyses. After this quality control step, 252 samples (see Supplementary Table 2) qualified for further analyses. In order to select a reference genome from each cluster (defined above), high-quality reads from a subset of metagenomes (TS1, TS4, TS25, MH0006, MH0012) were mapped to the 1,511 genomes with an alignment identity cutoff of 85% using Mosaik (version 1.1.0021; http://bioinformatics.bc.edu/marthlab/Mosaik) with the options "`-a all -m all -hs 15 -mmp 0.85 -mmal -minp 0.9 -mhp 100 -act 20`". Then, for each 95% genome cluster (see above), the genome with the highest read coverage was selected resulting in a set of 929 reference genomes, each likely representing a unique species (see Supplementary Table 2). Subsequently, all metagenomic sequence data sets (252 samples with minimum and average read length of 45 bp and 79.5 +/- 15.2 bp, respectively; see Supplementary Table 1) were mapped to these 929 reference genomes with an alignment identity cutoff of 95% using the same options as above, except for using "`-mmp 0.95`" instead of "`-mmp 0.85`". The number of reads mapping to a reference genome were counted and normalized by the genome size in order to obtain quantitative relative abundances of each genome in every sample.

1.3. Detecting the presence of a species in a sample

A recent study reported that two strains of *E. coli* are expected to share 40% of their genes[11]. Since complete genomes of several diverse strains of *E. coli* were used in that study, this can serve as a lower bound of the shared gene content for an average species. Therefore, to ensure that the reads from a gut sample are indeed derived from the same species as the reference genome they map to, we required that more than 40% of the nucleotide positions of that genome (breadth of genome coverage) are covered by reads from that sample – we then considered this species "present" in the sample. This procedure eliminated spurious cross-species mapping of reads in highly conserved genes. For example, when two species (A and B) are phylogenetically close to each other, highly conserved genes may be more than 95% identical between them. If a sample contains only species A, reads belonging to a highly conserved gene from species A will map to the region in species B containing the same gene, leading us to believe that species B was present in the sample. However, since species B is not present in the sample, this will likely not be enough to cover 40% of the genome of species B. Therefore all the spurious mappings of reads from species A to genome of species B will thus be ignored.

1.4 Prevalent and dominant species in our cohort

We identified 101 species, out of the 929 species in our database, which (i) are present in at least one sample (according to the detection criterion in Section 1.3) and (ii) accumulate a depth of genome coverage of at least 10x when summing over all samples. The second criterion avoids rare and transient species that are present in a small number of individuals. We consider these 101 to be the "prevalent" species in our cohort (Supplementary Table 2). 99% of the 7.4 billion reads mapped to just 66 out of the 101 species, suggesting that these are the "dominant" species in our cohort (Supplementary Table 2). For the prevalent species we observe on the pooled level more than 87% of genome coverage in more than half of the cases, and more than 56% in 90% of the cases (respectively, 68% and 45% on the individual sample level).

1.5. SNP calling

Multi-sample SNP calling was performed on the pooled samples. Bases with a quality of less than 15 were not considered. Single nucleotide variants were considered as SNPs if they had an allele frequency of at least 1%, which is the classical definition of polymorphism[12], and were supported by at least 4 reads. While the second criterion filters out sequencing errors randomly distributed across the genomes, the 1% criterion eliminates random sequencing errors that accumulate in the same position when depth of coverage reaches very high numbers. The resulting catalog of polymorphic sites contained 10.3 million entries.

We used our own custom scripts to perform multi-sample calling akin to the procedure outlined in Broad's GATK paper. The error estimates for the called variations are conservative false positive estimates in the range of 0.35-0.7%, which we believe, is entirely in support for the validity of our methodology. The use of customized calling procedures further allowed us to better utilize our computing infrastructure, collect additional statistics and derived data during the SNP calling process.

1.6. Structural variant and short insertion/deletion detection

Structural variants (SV; 50bp and larger) and short insertions/deletions (indel; under 50bp) were called using the Pindel program[13]. We extracted all reads suitable for use in the Pindel program, namely reads which mapped to reference genomes with indels as well as reads mapped to a reference genome whose paired-end read could not be mapped to reference genomes. A minimum number of four reads and a minimum of 1% of all reads (i.e., from the pooled samples) were required to support a SV/indel. In order to compare the rate of SV/indel and SNP calls, the reads were downsampled using the same methodology as used for SNP calling.

1.7. Estimation of error rates in SNP and SV/indel calling

To validate our SNP-calling procedure, we used two different approaches to calculate error rates in 40 essential single-copy marker genes[2,3] and extrapolated the results to the set of 101 prevalent species. First, we assumed that nonsense mutations in an essential gene represent errors and found 67 of those in 116,605 SNPs (0.06%) across 3,609 marker genes (Supplementary Table 4). However, mapping the position of these mutations revealed that 19 and 21 (60%) of these mutations occurred in the first and last 10 % of the marker gene lengths, respectively (Supplementary Fig. 2). A manual inspection of nonsense mutations within the first 10% of gene lengths revealed that start codon mis-predictions can partly explain this regional overrepresentation, while those in the last 10% may be due to gene length variations at the C-terminus, i.e., they likely do not represent premature but real stop codons. A linear regression of the cumulative number of nonsense mutations along the 10^{th} to 90^{th} percentile of gene length yielded a total of 32 (0.03%) expected errors. However, since there were likely other erroneous SNPs that did not cause stop codons, this number was to be viewed as a lower bound estimate. To estimate an upper-bound false positive rate, we calculated the codon frequencies for the marker genes in the 101 prevalent species and calculated a weighted probability of 4.13 % for stop codon producing SNPs (Supplementary Table 5). By extrapolating the probability-normalized error rate calculated for these marker genes to the whole set of 101 prevalent genomes (i.e., by dividing by the fraction of the genome length of the 101 prevalent species that is made up by the marker genes (1.06 %)), we calculated an upper bound false

positive rate 0.71% (when considering the accumulation rate of nonsense mutations in the 10 to 90 percentile of gene lengths), or 1.49% when considering all nonsense mutations (see formula below).

$$32 stop\_SNPs\_mg * \frac{100}{4.13} * \frac{100}{1.06} + 10.3 * 10^6 SNPs = 0.71\%$$

$$64 stop\_SNPs\_mg * \frac{100}{4.13} * \frac{100}{1.06} + 10.3 * 10^6 SNPs = 0.71\%$$

We performed a similar analysis to estimate an error rate for SVs and short indels by assuming that SVs and short indels resulting in frame shift mutations are erroneous. We detected no frame-shifting SVs and 79 frame-shifting short indels in the 3,609 marker genes of the 101 prevalent species. Again, it should be noted, that 68 of these indels occurred in the first or last 10 % of the gene lengths (Supplementary Figure 2), suggesting that many of these frame-shifts (in particular those in the last 10%) were not erroneous (see above). A linear regression of the cumulative number of stop-SNPs along the 10^{th} and 90^{th} percentile of gene length yielded a total of 8 frame-shift mutations. Extrapolation of this error rate, using the same method as for SNPs , to the total number of detected short indels (107,991) resulted in an upper bound for the false positive rate of 0.7 % (considering the frame-shifting short indels accumulation rate in the 10 to 90 percentile of gene lengths) or 6.9% if we consider the frame-shifting short indels over the whole gene length (Supplementary Figure 2)

$$8\ frameshift\_indels\_MG * \frac{100}{1.06} + 107991\ indels = 0.7\%$$

$$79\ frameshift\_indels\_MG * \frac{100}{1.06} + 107991\ indels = 6.9\%$$

Due to the novelty of our method, we used a second approach to estimate an error rate by assuming that missense mutation-causing SNPs at highly conserved positions in essential marker gene alignments are erroneous. We selected two genera, *Bacteroides* (36 genomes) and *Ruminococcus* (14 genomes), for which 11 and 10 genomes were selected, respectively, according to our generation of a non-redundant set of reference genomes. We constructed multiple protein sequence alignments of the marker genes in each genus, then for each reference genome, we mapped the SNP positions to the codon alignments and calculated the number of SNPs located at conserved sites (Supplementary Table 5). This procedure was performed twice using two different alignment methods[14,15] and although the alignments were slightly different, they yielded the same number of SNPs occurring in conserved positions. Assuming that missense mutations at conserved sites are errors yields a false positive rate of 0.35%.

1.8. pN/pS calculation

The expected ratio of non-synonymous and synonymous substitutions was calculated for detectable parts of genomes and genes. In order to calculate the expected ratio we assumed a uniform model for the occurrence of mutations across the genomic sequence. Depending on whether the ratio was to be calculated for all the detected parts of the genome or only a single gene we first took all the codons of the genome / gene and then observed the effect of all possible mutational events on the codon. The outcome of the mutational event was then classified as either synonymous of non-synonymous. Counting how often one of the two potential outcomes was observed resulted in the cumulative number of potential non-synonymous and synonymous mutational events (i.e. the expected ratio under the assumption of mutational uniformity).
Codons containing polymorphic sites were then extracted and the alleles categorized either as non-synonymous or synonymous. This observed ratio between non-synonymous and synonymous substitutions was then compared to the expected ratio, resulting in a measure akin to the classical dN/dS ratio used as proxy for evolutionary pressure.
dN/dS ratios are not applicable in our study since polymorphic sites are not assignable to individual strains, consequently not allowing us to have haplotypes, which are a prerequisite when calculating dN/dS ratios. On the other hand, pN/pS can be derived from variations in a pool of sequences without haplotype assignment.

1.9. Downsampling

When comparing SNP patterns in a given genome between two samples, we downsampled reads from both samples to the same depth of genome coverage in order to remove biases due to differences in coverage. For downsampling a sample to a specific coverage we calculated the ratio of target coverage and actual coverage. Then for each position in the genome we enumerated the different bases coming from different reads and retained each base from the sequencing data with a probability equal to the previously calculated ratio. This results asymptotically in the same coverage across the genome as would have been observed by removing the same fraction of reads prior to mapping. SNPs called prior to downsampling (from the set of 10.3 million) were retained if at least one base supports the allele after downsampling, resulting in a set of downsampled SNPs and downsampled polymorphic sites.

1.10 Estimating π and F_{ST}

Nucleotide diversity π is used to measure the degree of polymorphism in a population and is commonly defined as the average number of difference between the same regions of genomic DNA sampled randomly from a sample (population).

In the light of missing haplotypes, population sizes and high coverage variance across a genome we used

$$\pi(S,G) = \frac{1}{|G|}\sum_{i=1}^{|G|}\sum_{B_1\epsilon\{ACTG\}}\sum_{B_2\epsilon\{ACTG\}\setminus B_1}\frac{x_{i,B_1}}{c_i}\frac{x_{i,B_2}}{c_i-1}$$

in order to estimate π in a metagenomic sample. Where S is the sample, G is the genome of interest, $|G|$ the size of the genome, $x_{i,Bj}$ the number of nucleotide B_j seen at position i and c_i the coverage at position i in the genome.

π between two samples follows naturally from the above definition:

$$\pi(S_1,S_2,G) = \frac{1}{|G|}\sum_{i=1}^{|G|}\sum_{B_1\epsilon\{ACTG\}}\sum_{B_2\epsilon\{ACTG\}\setminus B_1}\frac{x_{i,B_1,S_1}}{c_{i,S_1}}\frac{x_{i,B_2,S_2}}{c_{i,S_2}}$$

With $x_{i,Bj,Sk}$ the number of nucleotide B_j seen at position i in sample S_k and $c_{i,Sk}$ the coverage at position i in sample S_k in the genome. Our approach derives from the π estimator on NGS data proposed by Begun et al[16] with extensions in order to account for more than two alleles per site. The formula works by calculating the chance of randomly choosing two different alleles (sigma two and three) at a randomly chosen base pair in the genome (sigma one).

The fixation index, a measure of population differentiation, is calculated using the standard definition:

$$F_{ST}(S_1,S_2,G) = 1 - \frac{\pi_{within}}{\pi_{between}} = 1 - \frac{\big(\pi(S_1,G)+\pi(S_2,G)\big)/2}{\pi(S_1,S_2,G)}$$

F_{ST} values are usually contained in the interval [0,1] with values around zero indicating highly similar samples close to one indicating strong differentiation (i.e. separate populations). Negative values can theoretically occur and are often either rounded to 0 or interpreted as out-breeding.

1.11. Calculating the shared allele fraction

To compare the gut microbial strain populations between two samples, we estimated the fraction of shared alleles between two samples using SNP patterns. For each genome, we found the samples for which there was at least 10x sequence depth coverage and downsampled them to 10x. We then calculated for each sample pair the number of downsampled polymorphic sites that were still present in both samples (shared polymorphic sites). Among these sites, we counted the number of sites that exhibited a shared allele. We then estimated the shared allele fraction by dividing the number of shared alleles by the number of shared polymorphic sites.

1.12. Shared allele similarity score

Assuming that each site is independent as done by (Goldman, JME 1993) and (Yang, 2006), the event of sharing the alleles across multiple polymorphic sites can be modeled using the binomial distribution. A limitation of the shared allele fraction calculated in Section 1.11 is that sharing the same fraction of alleles is much more significant when there are a larger number of polymorphic sites, which it ignores. For example, sharing 90 alleles out of 100 sites could result more easily by chance compared to 900 out of 1,000. As a simple example, the probability of obtaining a specific number of heads in a sequence of tosses of a fair coin can be modeled using binomial distribution. In this case, the probability of getting at least 1 tail out of 2 tosses, P(X>1; n=2, p=0.5), is 0.75. However, the probability of getting at least 10 tails out of 20 tosses, P(X>10; n=20, p=0.5), is only 0.588, significantly lower than the previous case. In both cases, we are estimating the probability of obtaining a tail in at least 50% of the cases. But as we increase the number of events in the sequence, that probability gets lower. Therefore we developed a shared allele similarity score based on the binomial distribution to account for this fact. We first estimated the expected shared allele fraction (p), which is the average shared allele fraction across all individuals (p=0.84 in our dataset). For two samples that share x alleles out of n polymorphic sites, we estimated the logarithm of cumulative probability of all possible events that share more than x sites.

$$F(x;n,p) = P(X > x) = \sum_{k=x+1}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$s(n,p) = -\ln F(x;n,p)$$

The cumulative probability represents the probability of sharing more than x alleles out of n polymorphic sites purely by chance. It takes on a value of 0 when all the alleles are shared, and a value of 1 when none of the alleles is shared, and could serve as a distance measure. Therefore the negative logarithm of this cumulative probability can already serve as a similarity measure that is not bounded between 0 and 1. However, since this value spans over several orders of magnitude (10^{-18} to $3x10^5$ in our dataset), for visual purposes we derived the shared allele similarity score after a second log transformation and linear rescaling using the maximum/minimum observed values to fit in the interval [0,1].

1.13. Calculation of F_{ST} and shared allele similarity scores in individuals and across continents

The calculation of all-by-all pairwise comparisons of F_{ST} values and shared allele similarity scores was based on 49 dominant genomes that were present (according to the detection criteria of 40% breadth of genome coverage) in at least two samples with >10x base pair coverage. F_{ST} was calculated for each sample pair as described in Section 1.10. The shared allele similarity score was calculated as

described in Section 1.12 using these 49 genomes between all sample pairs. For each of the 88 samples from 43 individuals for which more than one sampling time point was available, using the F_{ST} values and shared allele similarity score, we identified the most similar sample among (a) the remaining 251 samples and (b) the remaining samples from the 42 other individuals (i.e., non-self samples).
For continental separation analyses, the most similar sample originating from the same continent and the most similar sample from a different continent were identified using the shared allele similarity score on the cumulative set of 49 prevalent genomes (see above and Supplementary Table 12). To perform the analysis for individual genomes, we only considered samples with more than 10x base pair coverage and genomes with more than 10 samples representing each continent. For eight genomes fulfilling these criteria (Supplementary Table 15), we calculated F_{ST} as described in Section 1.10 by pooling all samples with more than 10x coverage for both continents. The genome with the highest F_{ST} value (*Bacteroides coprocola*) was tested for continental separation by sampling 10 individuals from each continent 100 times and calculate within versus between continent similarities using adjusted pairwise Mann-Whitney U-tests tests. Similarly, Mann-Whitney U-tests were performed on sub-sampled and 100 times bootstrapped data to test for differences between best self and best non-self similarity scores (47 best self vs. 47 sub-sampled (out of 94) best non-self scores) as well as for differences between continents. The breakdown by sample for the only genome that exhibited significant separation by continent, *Bacteroides coprocola*, is shown in Supplementary Table 14.

1.14. Genes exhibiting the highest and lowest pN/pS ratios

The pN/pS ratio was estimated for each of the 229,692 genes in the 66 dominant genomes independently for each sample. To reliably estimate pN/pS we required an average gene base pair coverage of 3 reads and discarded non-protein-coding genes. We then estimated the average pN/pS ratio of a gene across all samples. In order to ensure that we examined ubiquitous genes, we only considered genes that have a pN/pS ratio in at least half of the samples (>=126). In addition, where possible, genes were assigned to eggNOG v.2 Orthologous Groups (OGs; initially 8,122 OGs and 1,315 OGs after filtering steps) using blastp (bit score >60 bits) assuming orthology. The average pN/pS of each OG across 252 samples and the 66 dominant species was determined, by calculating the mean of the across-sample average pN/pS ratio for all genes in the OG. The OGs with the lowest and highest scores were analyzed.

1.15. *Eubacterium eligens* and *Roseburia intestinalis*

To investigate how different gut species respond differently to the pressure from the gut environment, we identified two genomes, which differed considerably in their pN/pS ratios (*R. intestinalis* – 0.236, *E. eligens* – 0.131), yet had similar

genome coverage (*R. intestinalis* - average coverage: 5.05x, sum coverage overall samples: 1,046x; E. eligens – average coverage: 5.65x, sum coverage overall samples: 1,169x) and were observed in similar number of samples (106 vs. 147, respectively). To reliably estimate pN/pS for each of the genes in a sample, we required that at least 3 reads from that sample covered each position in the gene and only considered protein-coding genes. We then estimated the median pN/pS ratio of each gene in these two genomes across samples. Mann-Whitney U-tests were performed to test for significant differences between the pN/pS spectrum of the two species. Orthologous groups (OGs) between the two species were identified using the eggNOG annotations. For each OG, the median across all samples was determined. For the common OGs between the two genomes, the log2 ratio between the OGs of the *E. eligens* (low pN/pS) and *R. intestinalis* (high pN/pS) was calculated.

To illustrate (i) the difference in the synonymous and non-synonymous mutation profile between genes with high and low pN/pS ratios and (ii) that the pN/pS ratios are not correlated with the SNP density, we identified a gene with high pN/pS in *R. intestinalis* with low SNP density (hypothetical gene ROI_05970, Fig. 2d), and a gene with low pN/pS in *E. eligens* with high SNP density (2-dehydro-3-deoxygluconokinase gene, Fig. 2e), in the same environmental conditions (both from sample 809635352-stool1). *E. eligens* gene has a higher polymorphism rate and is enriched in synonymous changes, while the *R. intestinalis* gene has a lower polymorphism rate and is enriched in non-synonymous changes.

2. Supplementary Notes

2.1. Sample origin

We used 252 samples originating from three sources. The MetaHIT project[7], NIH Human Microbiome Project[17] and Washington University[9]. The 110 MetaHIT samples from 110 individuals came from two cohorts from Denmark (71 samples) and Spain (39 samples). The Danish individuals were part of an obesity study cohort, where 42% of them are obese. The Spanish individuals are part of an Inflammatory Bowel Disease (IBD) study, where 64% have IBD. The 94 the NIH Human Microbiome Project samples consisted of 51 US-American individuals and were all considered “normal” as defined in the HMP sampling policy[17]. Of those 41 of the 51 individuals were sampled twice and 2 three times. The three Washington University samples came from three US-American individuals.

2.2. Motivation for a non-redundant set of reference genomes

Some microbial species (i.e., representative strains) have been sequenced more than others, which introduces a bias in the amount of variation that can be

detected in different species when metagenomic reads are mapped to a redundant set (i.e., several strains represent the same species) of genome sequences. For example, reads from a given species are more likely to match perfectly to a sequence if several genomes are available for that species, than if that species is represented by a single reference genome sequence only. As a consequence, this would bias the likelihood to identify polymorphic sites towards species with smaller numbers of reference sequences. To account for this bias, we generated a set of non-redundant reference genomes based on a median DNA sequence identity threshold of 95% in 40 single-copy marker genes[2,3]. With this non-redundant set, more than 91% of the reads mapped uniquely to a single genome each, despite the presence of highly conserved regions across reference genomes.

2.3. *galK* in *E. eligens* and *R. intestinalis*

Galactokinase (*galK*) is the first enzyme in the Leloir pathway, the only way to convert galactose into glucose[18]. This pathway is important for the usage of galactose and other galactose-containing substrates, like melibiose. While in *E. eligens* the median pN/pS of *galK* across 9 samples is 0.48, in *R. intestinalis* it is 0.03 across 7 samples. *E. eligens* cannot ferment galactose[19] suggesting that the Leloir pathway and *galK* are both non-functional. This is supported by experiments in *Bacillus subtilis* where it has been shown that for this pathway to be inactivated, galK must be non-functional, otherwise it would accumulate galactose derivative toxic for the cell[20]. On the contrary, *R. intestinalis* is able to use melibiose as a substrate[21], which seems to be dependent on *galK* activity. In support, *Thermus thermophilus* is only able to ferment melibiose in GAL $^{+}$ conditions (able to ferment galactose), suggesting that galactose accumulation inhibits microbial growth[22].

2.4. OGs with high SNP densities

In addition to OGs associated with conjugal transfer of antibiotics resistance there were two OGs representing CRISPR-associated proteins (Csy1 and Cse2), and many OGs lacking annotations (comprised of hypothetical proteins). Conjugation transfers a single strand of DNA and is highly error-prone, thus producing higher genomic variation. The fact that antibiotic resistance factors are carried on conjugative transposons implies that bacteria are able to rapidly evolve to evade new antibiotics. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) represent a new RNA-based immune system, which provides bacteria with sequence-specific resistance against phage, and other invading mobile genetic elements. The CRISPR spacers and phage genomes are constantly evolving to evade each other, and therefore the CRISPR loci may have high SNP density due to the high mutational pressures. However, the proteins involved in the CRISPR-system are not well characterized. In general the CRISPR-associated proteins are involved in processing RNA and DNA and include helicases and endoribonucleases[25,26]. Some of these proteins (including Csy1[23]) operate in

complexes and have sequence- and structure-specific interactions with CRISPR RNA or the template DNA. In addition, the structure of Cse2 suggests that it interacts with nucleic acids[27]. It is possible that these proteins must maintain genetic diversity to provide specific interactions with rapidly changing CRISPR-RNA or phage-DNA sequences.

2.5. Data availability

The catalogue of polymorphic sites and the called indels and SVs can be found at http://vm-lux.embl.de/~schloiss/variation/supplement.tar.gz as a downloadable package. Details on the files and formats are included (README.txt).

3. References

1 Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucl. Acids Res.* **38**, D190-195, doi:10.1093/nar/gkp951 (2010).

2 Ciccarelli, F. D. *et al.* Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* **311**, 1283-1287, doi:10.1126/science.1123061 (2006).

3 Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. *Science* **318**, 1449-1452, doi:10.1126/science.1147112 (2007).

4 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

5 Gish, W. (1996-2008).

6 Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10**, 504-509, doi:10.1016/j.mib.2007.08.006 (2007).

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

8 Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Research* **19**, 2317-2323, doi:10.1101/gr.096651.109 (2009).

9 Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).

10 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-180, doi:10.1038/nature09944 (2011).

11 Touchon, M. *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genetics* **5**, e1000344, doi:10.1371/journal.pgen.1000344 (2009).

12 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

13 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized

insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).
14 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:msb201175 [pii] 10.1038/msb.2011.75 (2011).
15 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh34032/5/1792 [pii] (2004).
16 Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS biology* **5**, e310, doi:10.1371/journal.pbio.0050310 (2007).
17 The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215-221, doi:10.1038/nature11209 (2012).
18 Frey, P. a. The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **10**, 461-470 (1996).
19 Holdeman, L. V. & Moore, W. E. C. New Genus, Coprococcus, Twelve New Species, and Emended Descriptions of Four Previously Described Species of Bacteria from Human Feces. *International Journal of Systematic Bacteriology* **24**, 260-277, doi:10.1099/00207713-24-2-260 (1974).
20 Krispin, O. & Allmansberger, R. The Bacillus subtilis galE gene is essential in the presence of glucose and galactose. *Journal of bacteriology* **180**, 2265-2270 (1998).
21 Duncan, S. H., Hold, G. L., Barcenilla, A., Stewart, C. S. & Flint, H. J. Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *International journal of systematic and evolutionary microbiology* **52**, 1615-1620 (2002).
22 Fridjonsson, O. & Mattes, R. Production of recombinant alpha-galactosidases in Thermus thermophilus. *Applied and environmental microbiology* **67**, 4192-4198, doi:10.1128/aem.67.9.4192-4198.2001 (2001).
23 Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10092-10097, doi:1102716108 [pii] 10.1073/pnas.1102716108 (2011).
24 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).
25 Ebihara, A. *et al.* Crystal structure of hypothetical protein TTHB192 from Thermus thermophilus HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Science* **15**, 1494-1499, doi:10.1110/ps.062131106 (2006).
26 Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *The Journal of Biological Chemistry* **283,** 20361-20371, doi:10.1074/jbc.M803225200 (2008).

27 Agari, Y. *et al.* X-ray crystal structure of a CRISPR-associated protein, Cse2, from Thermus thermophilus HB8. *Proteins* **73,** 1063-1067, doi:10.1002/prot.22224 (2008).

4. Supplementary Tables

Supplementary Table 1. **Information on samples used in this study.** 252 samples are summarized according to their names, data source, subject ID, sampling time point, continent of origin (NA: North America; EU: Europe). Statistics on amount of high-quality sequencing reads/bases, average read length, and mapping rate to the non-redundant set of reference genomes are shown. The maximum mapping rate in a single sample was 75.1%, the minimum rate 11.2%.

Supplementary Table 2. **Information on reference genomes used in this study**. NCBI taxonomic identifiers and species/strain names for 1,497 (no reads from any sample mapped to 14 genomes of the 1,511) prokaryotic genomes are shown (first two columns). Genomes that were selected as references for a cluster of genomes (N_{total} = 929) are shown in boldface. Entries in the third column denote whether a genome was among the 101 prevalent, 66 dominant (these are included in the prevalent genomes), or the remaining non-prevalent genomes. Genomes that were not selected but are represented by a selected genome are defined as clustered. For details on the generation of a reference genome set see Supplementary Information.

Supplementary Table 3. **Statistics of read mappings to reference genomes**. A total of 7.4 billion (out of 17.85 billion; 41.6%) reads originating from 252 metagenomic samples were mapped to the set of 929 reference genomes (see Supplementary Table 2). The reference genomes are sorted by decreasing numbers of mapped reads across all 252 samples. These numbers (reads in all samples) are incrementally summed up and reported as percentages of the total number of mapped reads (cumulative percentage). The last row reports the number of unmapped reads.

Supplementary Table 4. **Stop codon-causing SNPs in marker genes.** The number of stop codon-causing SNPs in essential single-copy marker genes for 101 prevalent species used for estimating the false discovery rate of SNP calling.

Supplementary Table 5. **SNPs in conserved sites of marker genes.** The number of SNPs that occur in conserved sites of multiple sequence alignments of essential

single-copy marker genes for genomes belonging to the *Bacteroides* and *Ruminococcus* genera used for estimating the false positive rate of SNP calling.

Supplementary Table 6. **Sample polymorphic sites**. We identified in the set of 101 prevalent genomes (329 Mb positions in total) across all samples a total of 10.3 million SNPs and estimate an average nucleotide polymorphism rate of 3.1%. Within an individual, the rate is much lower, ranging between 0.68% and 1.77%.

Supplementary Table 7. **Orthologous groups with the highest SNP density.** We identified the 60 orthologous groups (OGs) with the highest SNP densities that occur in more than one third of the samples. The SNPs/kb was calculated by averaging across all genes in the same OG for all individuals (252 samples). This table also indicates the number of distinct proteins that occur in each OG.

Supplementary Table 8. **Matrix of Species-Sample average pN/pS across all genes.** pN/pS ratios for each species-sample pair of the 66 species and 207 samples (97 Americans - first time point only, 39 Spanish and 71 Danish) with at least 10x sequencing coverage have been calculated. Whole dataset pN/pS ratios have been calculated for the pooled samples and by averaging the sample ratios. NA indicates samples in which the genome had insufficient coverage to calculate a pN/pS ratio.

Supplementary Table 9. **Median pN/pS values of genes in *E. eligens* and *R. intestinalis***. For a total of 1,153 in *E. eligens* and 1,917 genes in *R. intestinalis* median pN/pS ratios were calculated across 207 samples (97 Americans - first time point only, 39 Spanish and 71 Danish). A total of 611 orthologous groups (OGs) were found in common between the two species, among them COG0153 (Galactokinase) has the highest ratio and COG0591 (Na+/proline symporter) is one of the OGs with lowest ratios. The ProteinIDs correspond to NCBI gi numbers.

Supplementary Table 10. **List of orthologous groups showing consistently high and low pN/pS values.** We identified the 70 OGs with highest and lowest pN/pS values, by averaging for each orthologous group (OG) the pN/pS values across 252 samples and 66 species. Among the OGs with lowest pN/pS values are OGs containing the genes *rpoB* (COG0085) and the genes involved in type IV secretion systems (COG3451 and COG3505). Relative to the OGs with the highest pN/pS values, there are transposases (COG2801 and COG3328) and bile salt hydrolase (COG3049).

Supplementary Table 11. **List of non-annotated genes with low pN/pS values.** 14 genes of unknown function have been found to be among the gut microbial genes with lowest average pN/pS values (across samples), present in at least 126 samples. The ProteinIDs correspond to NCBI gi numbers.

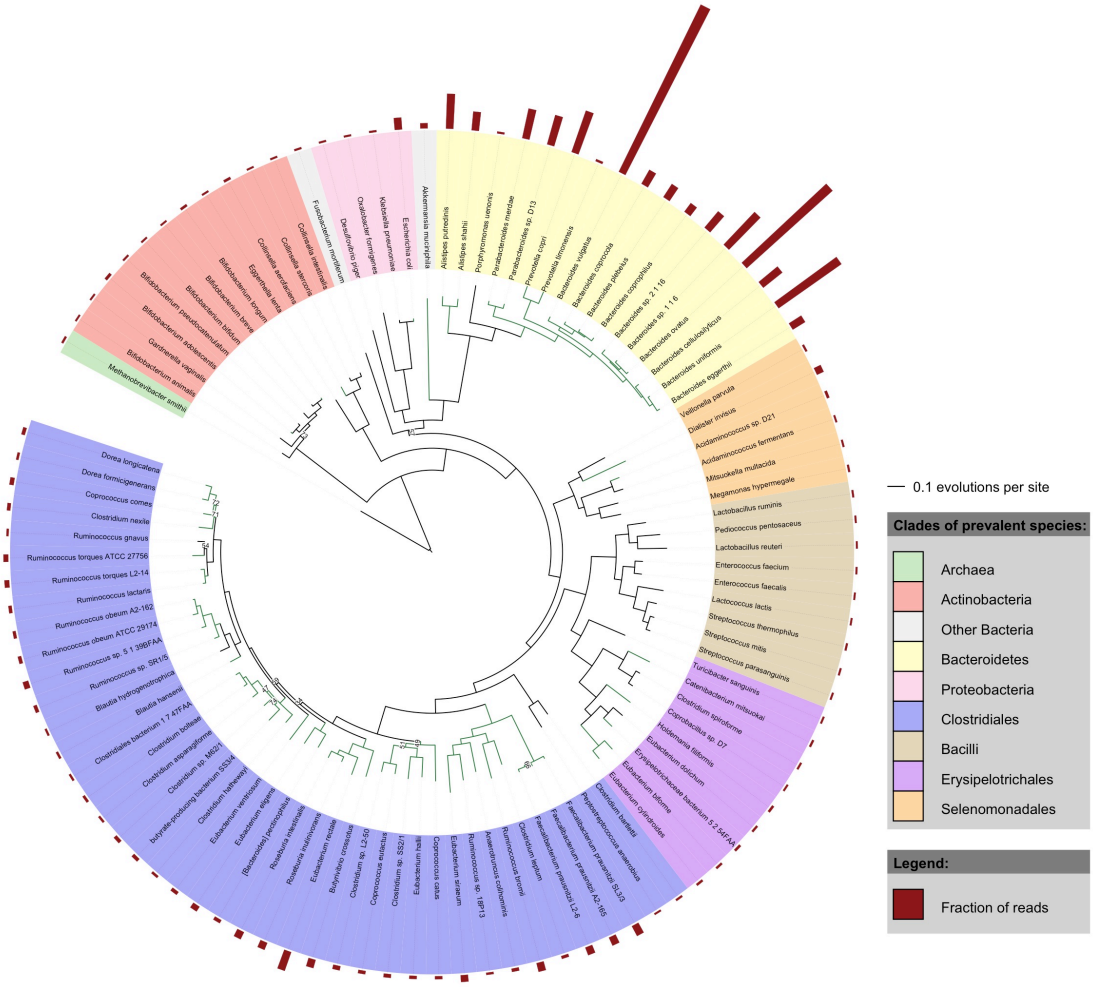
Supplementary Table 12. **Data summary for SNP sharing and taxonomic abundance similarities.** For 88 samples from 43 individuals that were sampled more than once (time-series data), time (in days) between sample points, scores for SNP sharing as well as the Jensen-Shannon Distance (JSD) similarities[10] for the most similar self and non-self samples are shown. Only one sample, 763536994-stool2 (highlighted in red), did not have the highest SNP similarity score to a sample that originated from a different time point from the same individual.

Supplementary Table 13. **Data summary for intra- and inter-individual SNP sharing by genome.** For 35 genomes for which intra-individual comparisons were possible (columns 1 and 2), the number of self (column 3) and total (column 4) pair-wise SNP sharing scores are shown with the number of incidences in which the score for an intra-individual comparison was higher than all inter-individual comparisons (columns 5 and 6).

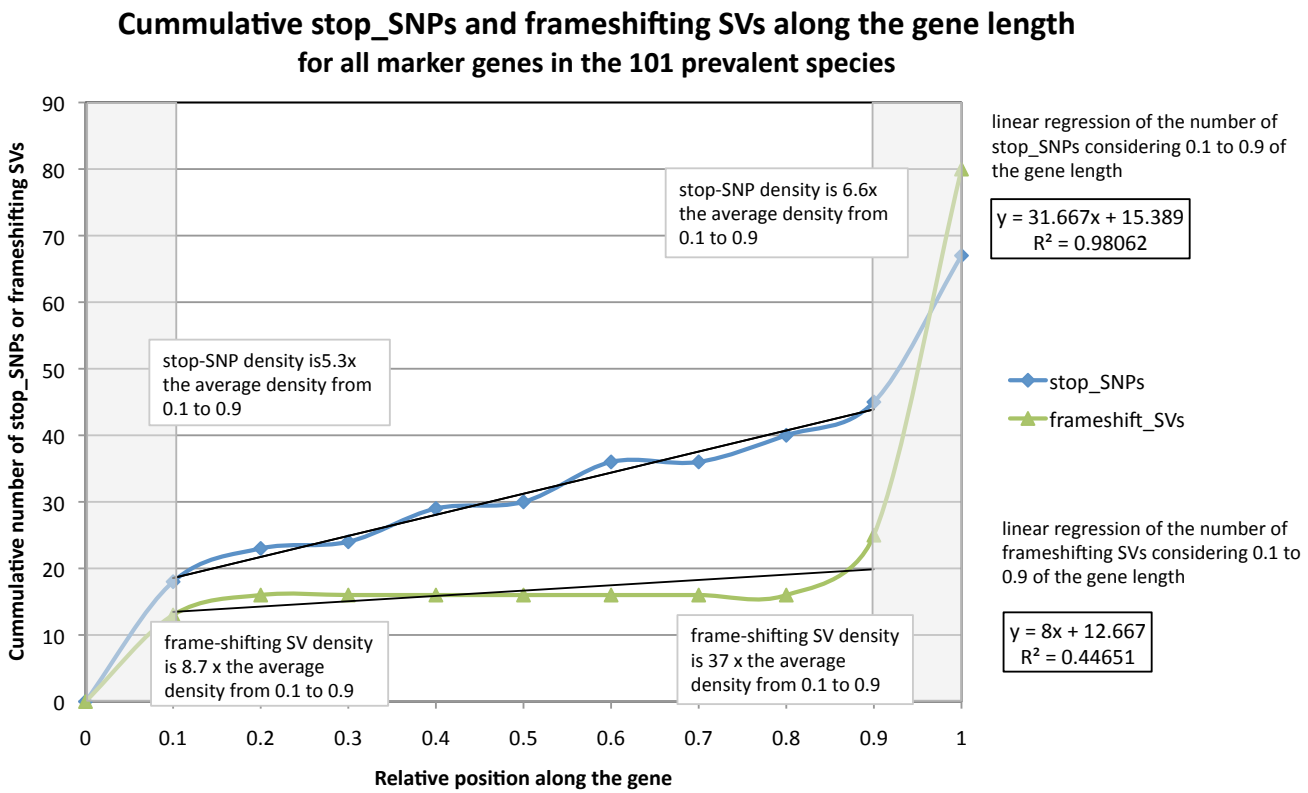
Supplementary Table 14. **Continental separation of SNP variation patterns.** SNP similarity scores between gut microbial samples and the most similar sample from the same continent (Similarity best hit same continent) and the most similar sample from a different continent (Similarity best hit different continent) were calculated as well as the difference between these scores (Difference best hits). This analysis was performed using all polymorphic positions identified on 49 dominant genomes (see Supplementary Methods 1.12), or on each of the 49 genomes separately (data for *B. coprocola* are shown; for analysis by genome see Supplementary Table 15).

Supplementary Table 15. **Continental separation analysis on single genomes.** We required a minimum of 10 samples from each of the two continents (NA: North America and EU: Europe) for between continent comparisons. For 8 genomes, within and between continent F_{ST} and the median allele sharing score for within and inter-continental sample pairs (i.e., North American/European or European/North American) was calculated. Only for one (*B. coprocola*) we found evidence for intercontinental differences.

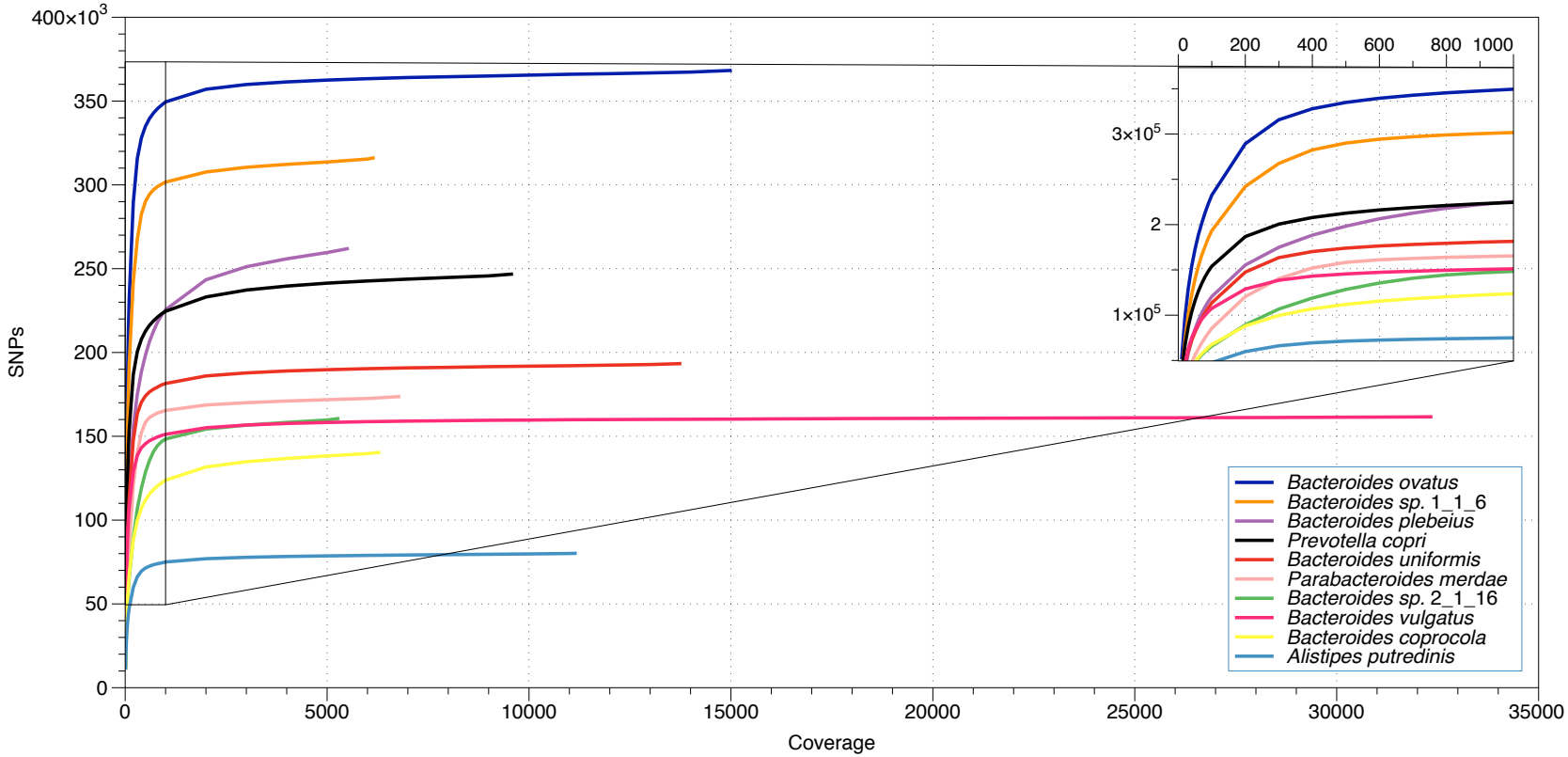
5. Supplementary Figures



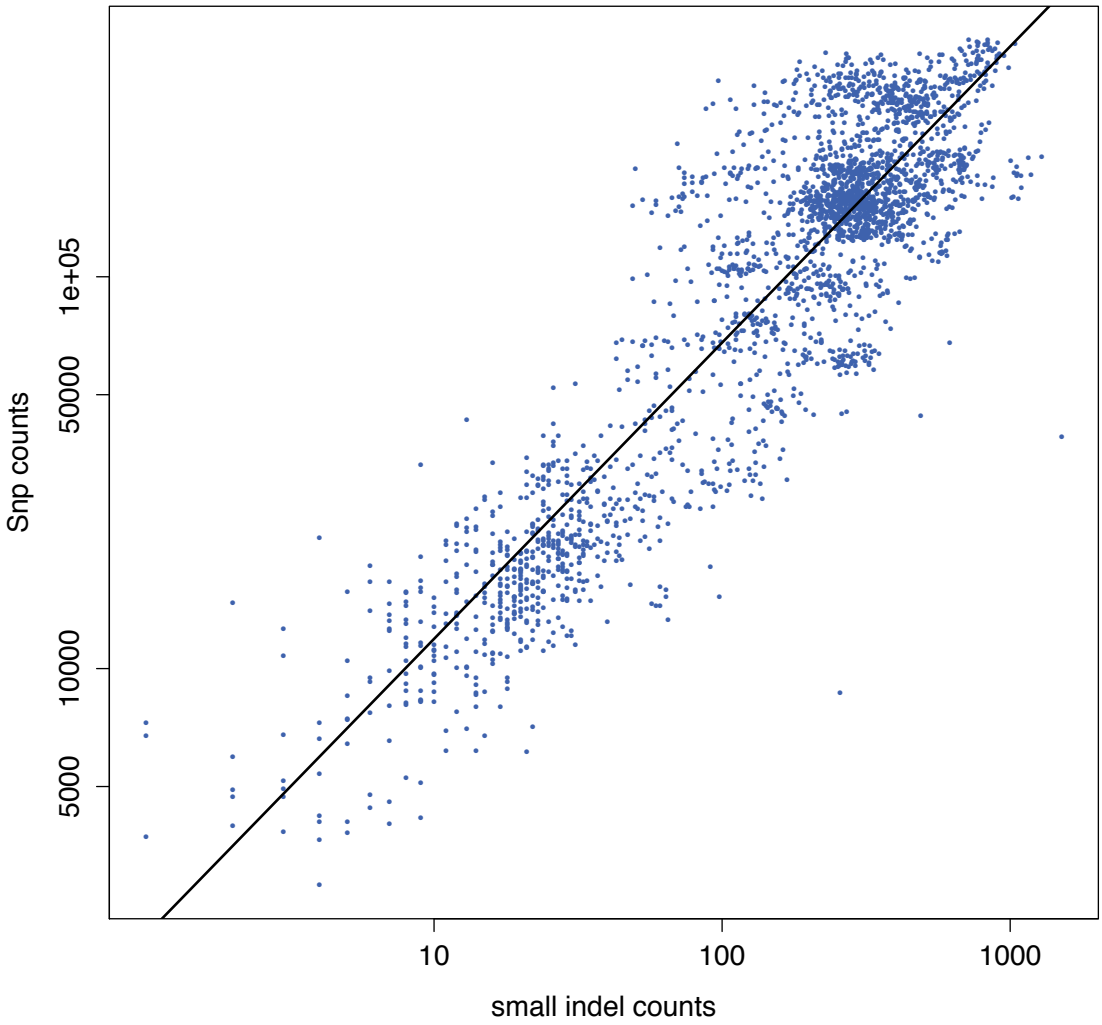
Supplementary Figure 1. **Phylogenetic tree of prevalent and dominant species in the combined MetaHIT, HMP cohort.** 101 prevalent species including the 66 dominant species (marked by green branches) are shown in a maximum likelihood phylogenetic tree. The 66 dominant species account for more than 99% of the reads mapped to the 929 reference genomes. Due to the difference in genome size these 66 species don't coincide with the species that have the highest base pair coverage. The fraction of reads mapped to each prevalent genome is shown as a bar chart. The tree was constructed from 40 marker genes[2,3] with RAxML v7.2.8 [24] using PROTGAMMAWAG model. Bootstrap proportion values from 100 replicates are shown only for branches with less than 80% support.



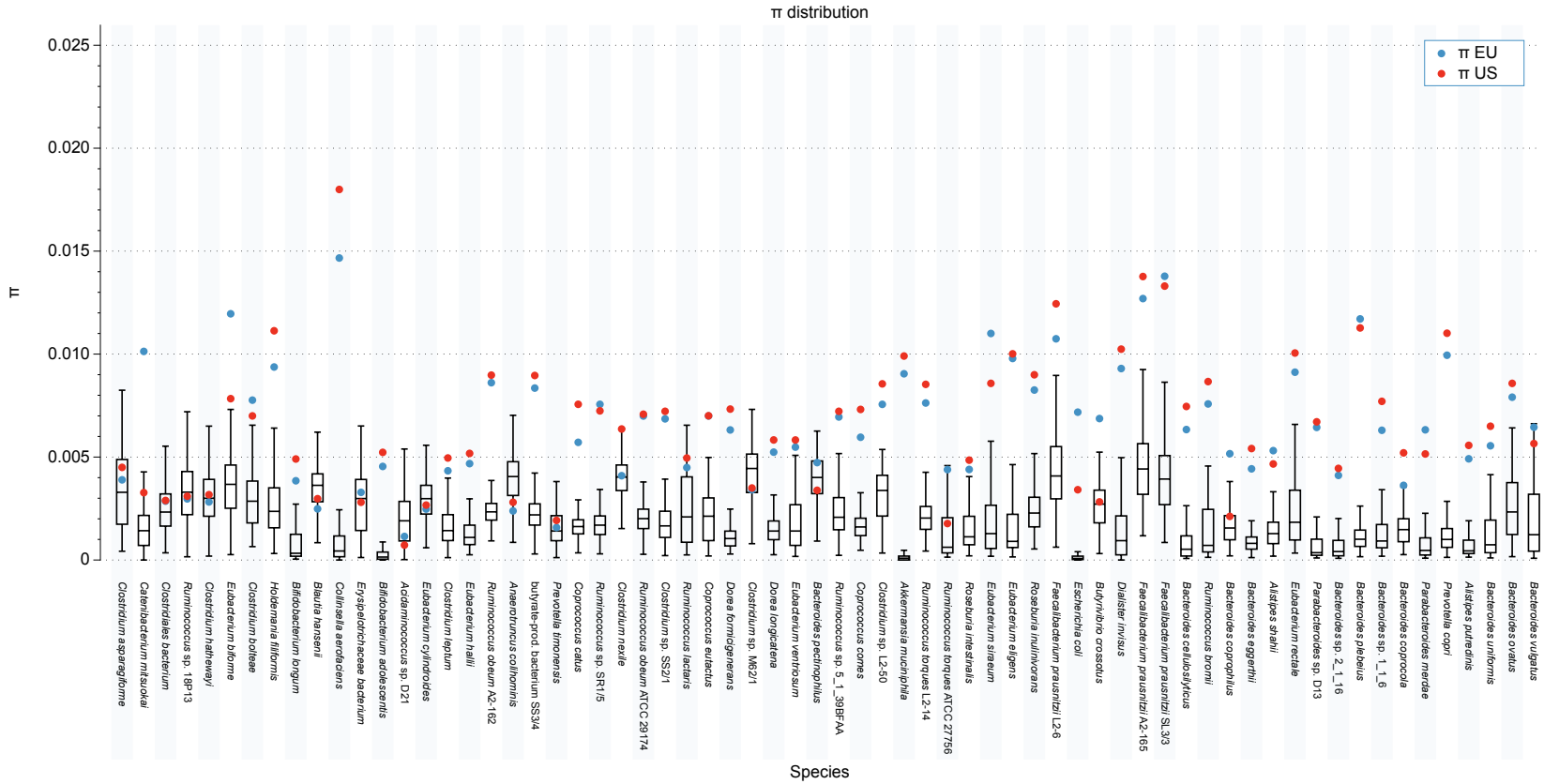
Supplementary Figure 2. **Cumulative stop-SNPs and frame-shifting SVs along the gene length for all marker genes in the 101 prevalent species.** This plot shows the cumulative number of SNPs that produce stop codons (blue) and structural variants that produce frameshifts (green), for each 10 percent of the proteins length. The grey shaded regions highlight that the stop-SNP density and frame-shifting density is much higher in the first and last 10 percent of the gene lengths. Manual inspection showed that many of these SNPs are due to improper annotation of gene start and stop positions.



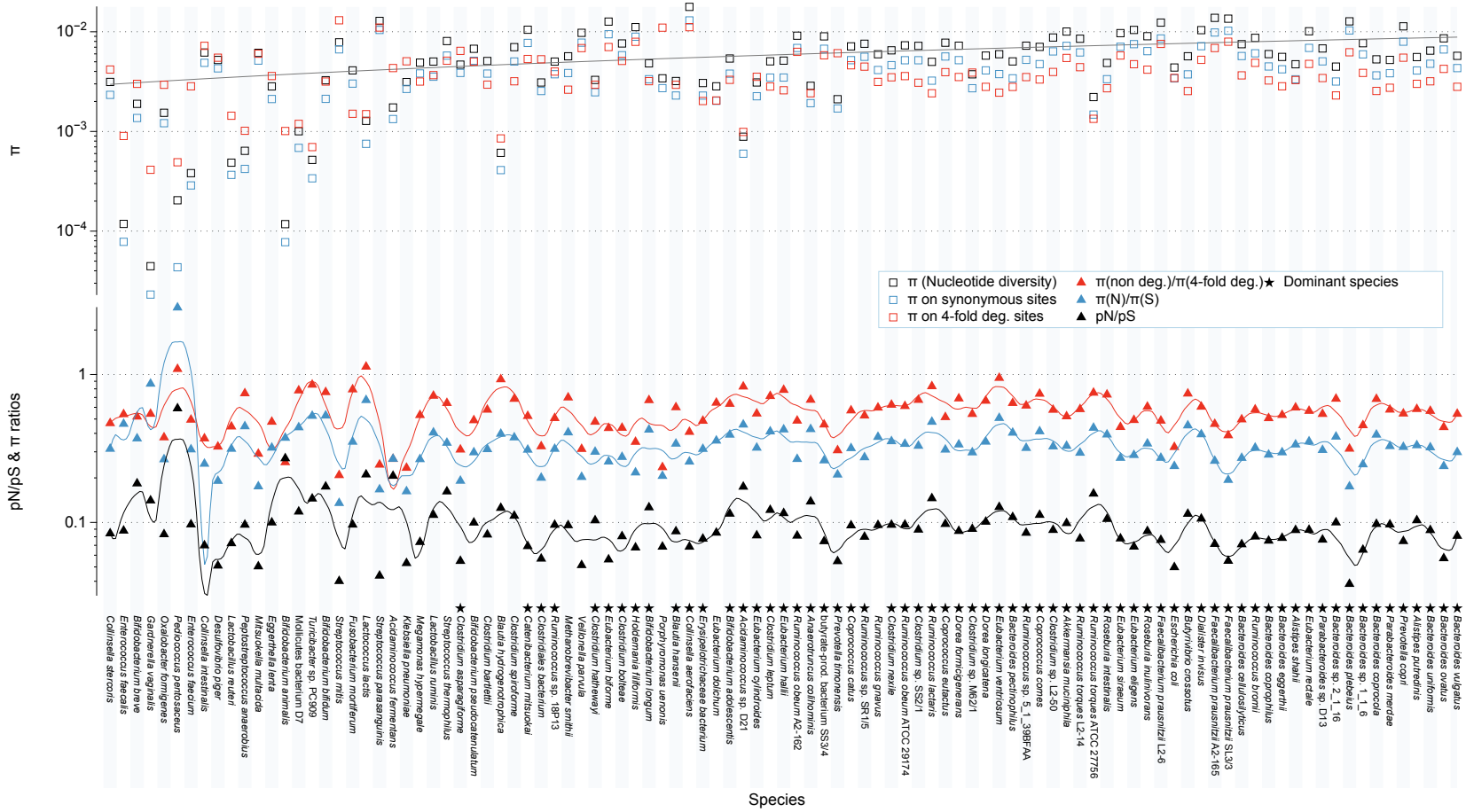
Supplementary Figure 3. **Downsampling of SNPs.** SNPs of the overall ten most abundant species were downsampled, starting from their native coverage down to 10x. The growth in the number of SNPs starts slowing considerable in the <1,000x range and basically stops afterwards, as only rarer alleles are discovered in the sequence data.



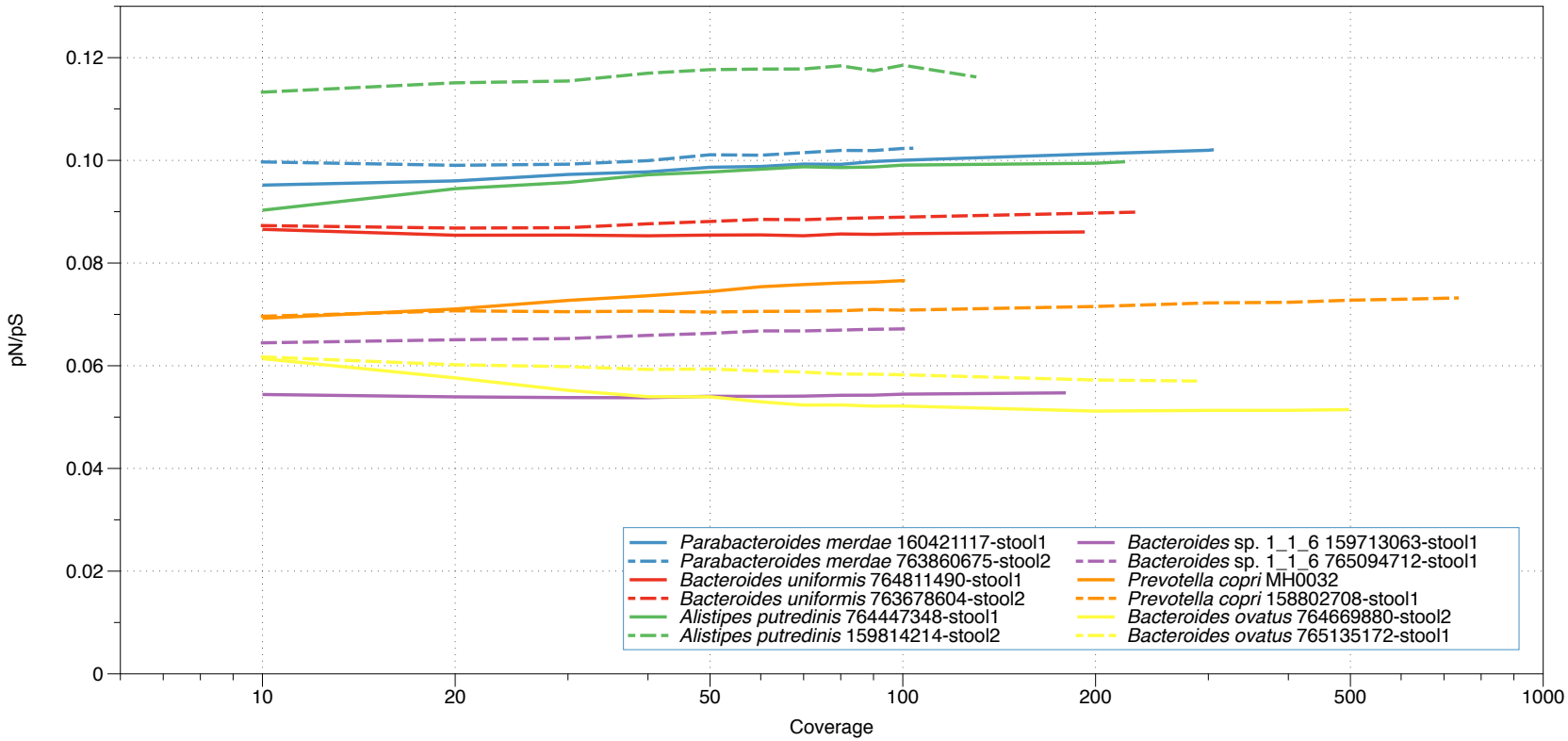
Supplementary Figure 4. **Dot plot of the logarithm of the number of SNPs vs. small indels (<50bp) for every sample-genome combination fulfilling the above-mentioned criteria.** Counts of SNPs and small indels were downsampled to 10x coverage. The Pearson correlation coefficient was 0.86 and the correlation was highly significant ($p < 2.2x10^{-16}$ - test statistic based on Pearson's product moment correlation coefficient.



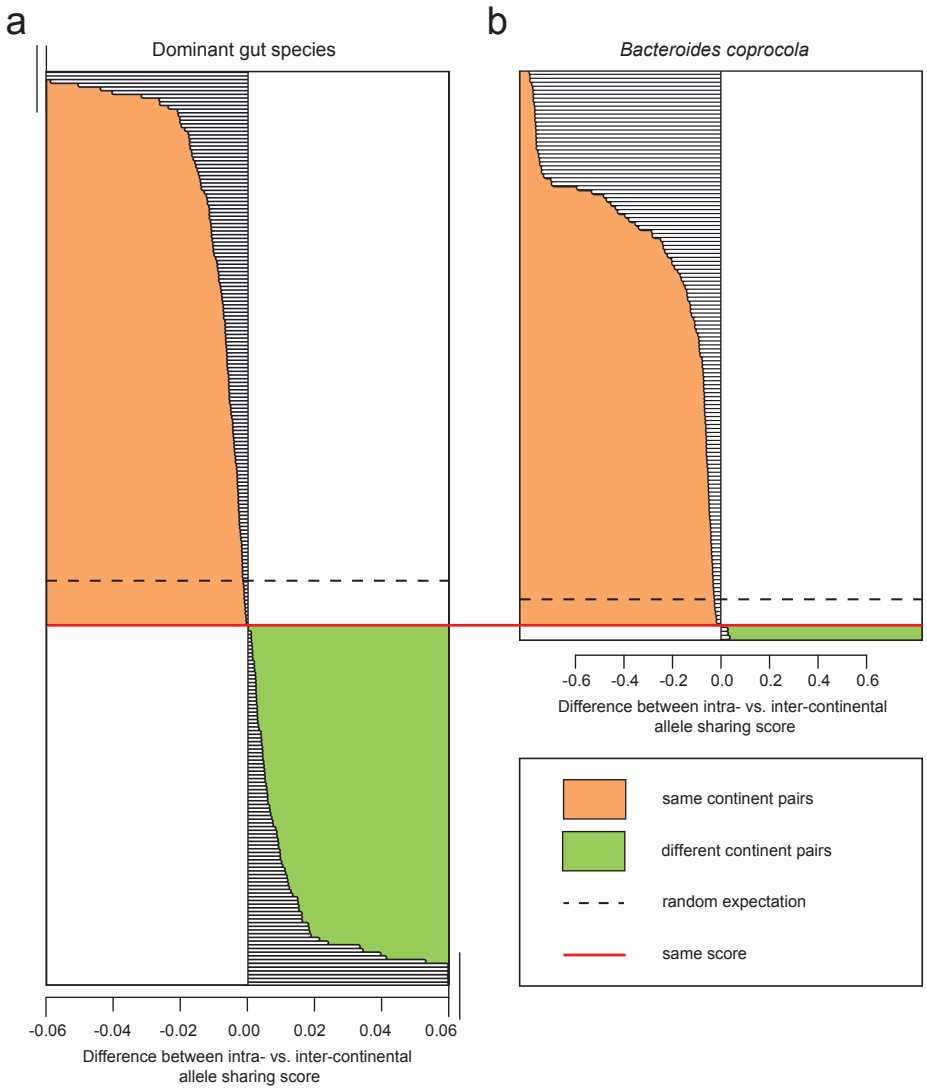
Supplementary Figure 5. **Distribution of π across all samples.** Box plots of the π distribution in the 66 dominant species. For comparison we show π values derived from the pooled European and North American samples as blue and red dots respectively.



Supplementary Figure 6. **π for different sites and its derived ratios.** π for synonymous, four-fold degenerate and all sites is shown. Derived ratios π(N)/π(S), π(non-degenerate)/π(four-fold degenerate) are plotted and highly similar to the pN/pS ratio.



Supplementary Figure 7. **Influence of downsampling on pN/pS.** SNPs of four dominant species, each from two samples, were downsampled, starting from their native coverage down to 10x. At each downsampling step the SNPs remaining after the downsampling were used for the calculation of the genome pN/pS ratio. The plot shows a largely stable ratio across the whole coverage range for the eight instances. In addition, for all genome/sample pairs with a coverage of at least 50x (635 pairs; in order to have a significant influence of the downsampling on the number of SNPs) we performed downsampling to 10x and for 87% percent of those (552 pairs) the difference between the pN/pS at 10x and the native coverage was less than 0.01.



Supplementary Figure 8**. Inter-continental comparison of gut microbial species.** Using the allele sharing similarity score (**a**), for each sample, the most similar sample from another individual (i) in the same continent and (ii) in a different continent were identified. For a given sample, the difference in the allele sharing scores (i)-(ii) was calculated, with positive and negative values indicating that allele sharing is stronger within the same continent and between different continents respectively. Samples (y-axes) are ordered by the difference in their scores, with positive differences in green and negative differences in orange. Scores from 49 dominant species (**a**) show that the number of positive and negative scores is not significantly different from random expectation and that the magnitude and distribution of the positive and negative scores are not significantly different, implying that there is no overall tendency for stronger allele sharing within the same continent. When estimating scores for eight dominant species individually (**b**), *Bacteroides coprocola* was the only species where positive scores were significantly higher in magnitude than negative scores (Supplementary Table 14 and 15) implying a significantly stronger allele sharing for *B. coprocola* within

individuals from the same continent. For differences ranging out of scale, (8 samples in (**a**)), bars were pruned to fit into the plot (all data available in Supplementary Table 14).