Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria

Stephen Nayfach[1,2], Katherine S. Pollard[1, 2, 3, 4 §]

1 Integrative Program in Quantitative Biology, University of California San Francisco, CA 94158
2 Gladstone Institutes, University of California San Francisco, CA 94158
3 Institute for Human Genetics, University of California San Francisco, CA 94158
4 Department of Epidemiology and Biostatistics, University of California San Francisco, CA 94158

§ Corresponding author

Email addresses:
SN: stephen.nayfach-battilana@ucsf.edu
KSP: kpollard@gladstone.ucsf.edu

Abstract

Deep sequencing has the potential to shed light on the functional and phylogenetic heterogeneity of microbial populations in the environment. Here we present *PhyloCNV*, an integrated computational pipeline for quantifying species abundance and strain-level genomic variation from shotgun metagenomes. Our method leverages a comprehensive database of >30,000 reference genomes which we accurately clustered into species groups using a panel of universal-single-copy genes. Given a shotgun metagenome, *PhyloCNV* will rapidly and automatically identify gene copy number variants and single-nucleotide variants present in abundant bacterial species. We applied *PhyloCNV* to >500 faecal metagenomes from the United States, Europe, China, Peru, and Tanzania and present the first global analysis of strain-level variation and biogeography in the human gut microbiome. On average there is 8.5x more nucleotide diversity of strains between different individuals than within individuals, with elevated strain-level diversity in hosts from Peru and Tanzania that live rural lifestyles. For many, but not all common gut species, a significant proportion of inter-sample strain-level genetic diversity is explained by host geography. *Eubacterium rectale*, for example, has a highly structured population that tracks with host country, while strains of *Bacteroides uniformis* and other species are structured independently of their hosts. Finally, we discovered that the gene content of some bacterial strains diverges at short evolutionary timescales during which few nucleotide variants accumulate. These findings shed light onto the recent evolutionary history of microbes in the human gut and highlight the extensive differences in the gene content of closely related bacterial strains. *PhyloCNV* is freely available at: **https://github.com/snayfach/PhyloCNV**.

Introduction

Microbial communities play a myriad of important roles in a number of ecosystems including marine, soil, and human-associated habitats. Often, microbial species are not isogenic organisms, but exist as complex populations of cells with heterogeneous genomes that confer different functions and ecology. For example, differences in gene content at the strain level can confer drug resistance [1, 2] or convert a commensal bacterium into a pathogen [3-5]. Therefore, quantifying strain-level variation within and between communities is critical for precision medicine. To use the microbiota for diagnosis, consider it in drug dosing, or curate it as a form of treatment, we need a more comprehensive understanding of how microbial genes vary across strains. More broadly, strain-level differences can shed light on recent evolution and adaptation of microbes in many environments, including the prevalence of ecological niche specialization and the extent to which different species have biogeography.

Metagenomic shotgun sequencing has the potential to shed light onto strain-level heterogeneity among bacterial genomes within and between microbial communities, yielding a resolution not achievable by sequencing the 16S ribosomal RNA marker gene alone [6, 7], and circumventing the need for culture-based approaches. However, limitations of existing computational methods have prevented most researchers from obtaining this level of resolution from metagenomic data. Assembly-free methods that estimate the relative abundance of known strains [8-10] may be effective for well-characterized pathogens like *E. coli*, but do not have power to detect strain-level variation for the vast majority of known species with only a single sequenced representative. Other approaches have been described for identifying single-nucleotide variants [11] and structural copy-number variants [12, 13] of microbial populations from metagenomes, but have not been integrated together or made available as software. Other methods [14, 15] use single-nucleotide polymorphism patterns to phylogenetically type strains, but do not capture the functions of these organisms and therefore cannot shed light onto the ecological forces shaping their genomes. Finally, assembly-based methods [16] often struggle with the presence of multiple closely related species present in a community and usually require manual inspection of assembled contigs.

To address these issues, we developed *PhyloCNV*, which is a computational pipeline that quantifies bacterial species abundance and strain-level genomic variation from shotgun metagenomes. Our method integrates many features that are absent from current software (Table S1) and leverages a comprehensive database of >30,000 reference genomes which we accurately clustered into species groups using a panel of universal-single-copy genes. Given a shotgun metagenome, *PhyloCNV* will rapidly and automatically identify gene copy number variants (CNVs) and single-nucleotide variants (SNVs) present in abundant bacterial species. By coupling fast taxonomic profiling via a panel of universal-single-copy genes with sensitive pan-genome and whole genome alignment, *PhyloCNV* can efficiently and automatically analyze hundreds of metagenomes. We applied *PhyloCNV* to faecal metagenomes of diverse human hosts with the goal of understanding global patterns of strain-level variation and biogeography in the human gut microbiome. We found that there is more nucleotide diversity of strains between different individuals than within individuals, that many species' population structure tracks with host geography, and that the gene content of some bacterial strains diverges at short evolutionary timescales in the human gut. *PhyloCNV* is implemented in Python and is freely available, along with documentation at: **https://github.com/snayfach/PhyloCNV**.

Results

A data resource for profiling strain-level variation

To quantify species abundance and strain-level genomic variation more broadly and accurately than existing approaches, we sought to compare metagenomic sequences to all known genomes. Towards this goal, we compiled a comprehensive database of 31,007 bacterial reference genomes [17]. We then systematically identified species groups among these genomes using an automated, sequence-based approach in order to avoid inconsistent, erroneous, and incomplete annotations that afflict some microbial taxonomies [18], and to greatly expand and improve upon previous efforts to systematically delineate bacterial species [11, 18-20]. Specifically, we hierarchically clustered reference genomes using the pairwise percent identity across a panel of 30 universal genes that we manually selected from a panel of 112 candidates [21] (Figure 1a, Table S2, and Methods). We found that the best genes for identifying bacterial species were less conserved and more widely distributed relative to other genes we tested (Figure S1). Using these marker genes increases computational throughput compared to whole genome comparisons, while producing genome-clusters that are highly concordant with a gold standard definition of prokaryotic species based on 95% genome-wide average nucleotide identity (ANI) [22, 23] (Table S3).

Our procedure clustered the 31,007 bacterial genomes into 5,952 genome-clusters with >95% ANI, representing distinct bacterial species (Tables S4-S6). When compared to a reference taxonomy (Figure 1b), we were able to classify 2,666 genomes (8.6% of total) that had not been previously annotated at the species level and to reassign species labels for 3,035 genomes (9.8% of total) that had been either assigned the incorrect species label based on the 95% ANI criterion (N=1,380) or were split from a larger cluster with the same species name (N=1,655). Previous work that implemented a similar genome-clustering procedure [18] found that the majority of disagreements with taxonomy could be supported by the literature.

We next leveraged these genome-clusters to compile a comprehensive genomic data resource for *PhyloCNV* (Figure 1c). First, we extracted a panel of 15 universal-single-copy genes from each genome-cluster (88K genes) to use for rapid metagenomic species profiling, which were selected based on their ability to accurately recruit metagenomic reads to the correct species (Methods). Second, we identified the set of unique genes within each genome-cluster to use for metagenomic pan-genome profiling (32M genes). By eliminating redundant genes with >99% nucleotide identity, we were able to reduce the number of reference genes by 73% without loss of functional diversity. This data compression facilitates more rapid analysis of strain-level variation. We further clustered these non-redundant genes at different levels of sequence identity (75-95% identity) in order to identify *de novo* gene families of varying size and diversity for downstream analyses. Functional annotations for these genes were obtained from PATRIC [17] and include FIGfams [24], Gene Ontology [25], and KEGG Pathways [26]. Third, we identified a representative genome from each cluster to use for identifying SNVs from metagenomes (5,952 genomes). The resulting database provides a foundation upon which we developed *PhyloCNV* and should be of value to other methods and studies in the future. All data resources are freely available and can be found online at: **http://lighthouse.ucsf.edu/phylocnv.**

An integrated pipeline for quantifying strain-level variation from shotgun metagenomes

Next, we developed *PhyloCNV,* which is an integrated software tool that processes shotgun metagenomes to sensitively and automatically quantify species abundance and strain-level CNVs and SNVs for any of the ~6,000 bacterial species in our database (Figure 2). CNV profiles offer an opportunity to quantify functional diversity across strains of each microbial species and to identify core genes that are consistently present in strains. SNVs in core genomic regions are useful for quantifying population diversity, dynamics, demography, and evolution. Each component of *PhyloCNV* was evaluated and optimized for speed and accuracy with simulations of 20 metagenomes created from real Illumina reads in order to capture sequencing error and biases that occur in real datasets (Tables S7-8 and Methods). Our software is freely available and can be found online at: **https://github.com/snayfach/PhyloCNV.**

In the first step of the pipeline, *PhyloCNV* estimates the coverage and relative abundance of species present in a metagenome by rapidly mapping reads to a database of phylogenetic marker genes (Table S9) and probabilistically assigning reads to species groups (Figure 2a and Methods). Relative abundances are measured in terms of the proportion of cells rather than the proportion of reads and thus can account for differences in average genome size and the fraction of unknown species between communities [27]. This taxonomic profiling step enables automatic quantification of strain-level genomic variation in downstream steps without any prior knowledge about a community's composition, and it avoids wasteful alignments to genes and genomes from sequenced organisms that are not present in the community. Existing methods for metagenomic taxonomic profiling that rely on taxonomic labels and are not up-to-date with currently sequenced genomes were not suitable for this step. *PhyloCNV* can accurately estimate abundance relative to other known species using relatively few reads (~1M reads), even though genome coverage of these species is slightly underestimated in simulations (Figure S2 and Table S10). The speed of *PhyloCNV's* taxonomic profiling (5000 reads/second) is comparable to other widely used methods, including MetaPhlan [28] and mOTU [7] (Figure S3).

After abundant species have been identified, *PhyloCNV* rapidly and sensitively quantifies gene copy number variation among strains of these species (Figure 2b and Methods). Instead of mapping reads to all reference genomes, *PhyloCNV* only maps reads to genes from the subset of abundant species, which maximizes throughput and minimizes memory consumption. Genes are determined to be present or absent based on their coverage relative to a panel of universal single-copy genes

(Figure 3b), which can be done accurately for strains with at least 3x sequencing coverage (Figure S2). We found that *PhyloCNV* is much faster than existing methods, such as PanPhlAn (http://segatalab.cibio.unitn.it/tools/panphlan), for quantifying CNVs from metagenomes, and the algorithm scales remarkably well with the number of species searched and the number of sequenced genomes per species (Figure S4).

In the final step of the pipeline, *PhyloCNV* identifies nucleotide variants among strains of abundant species by mapping reads to a database of representative genomes (Figure 2c and Methods). After read mapping, *PhyloCNV* estimates SNV allele frequencies, which enable estimation of strain-level heterogeneity within and between samples. Additionally, consensus sequences are used to build phylogenetic trees, which enable phylogenetic placements of strains relative to other samples and other reference genomes. Using metagenomic simulations, we found that we can consistently call consensus SNVs at a low false-discovery rate, although at least 5 to 10x coverage is required to identify the majority of SNVs (Figure S2). Filtering bases according to their mapping and sequence quality scores marginally improves precision while reducing the number of variants identified (Table S11). Additionally, we found that we can use our approach to accurately estimate the phylogenetic distance of novel strains to sequenced representatives (Figure S2).

Identification of globally distributed species in the human gut microbiome

We next applied *PhyloCNV* to >500 faecal metagenomes from diverse human populations with the goal of understanding global patterns of strain-level variation and biogeography. These data included samples from healthy unrelated individuals from Europe (Austria, Denmark, Spain, Sweden) [29-31], the United States [32], China [33], Tanzania [34], and Peru [35] (Table S12). To our knowledge, this is the first global-scale analysis of strain-level variation in the human gut microbiome.

Towards this goal, we first estimated the cellular relative abundances of bacterial species across these datasets and found that our reference database could account for between 30-60% of the species abundance in most gut communities (Figure 3a). This finding is consistent with a previous report [7] and indicates that a large proportion of bacterial genomes in the human gut have a sequenced representative at the species level. Strikingly, gut microbiomes of subjects from Tanzania and Peru have higher levels of novel species with no sequenced representative in our database, revealing strong bias of genome sequencing to date towards species associated with hosts in the northern hemisphere. This finding is supported by previous work, which found elevated levels of novel genera [36] and functions [34] in the gut microbiome of hunter-gatherers. In contrast to species, strains in human metagenomes are not well represented by sequenced genomes. We found that on average only 8% of strain-level phylogenetic diversity was shared with reference genomes, indicating that even for known, prevalent gut species there remains a tremendous amount of strain-level diversity in the human population that has not yet been captured by genome sequencing projects (Figure S5 and Methods).

We next performed principal component analysis on species relative abundances, and found that the variation between gut communities along the first three principal components (PCs) is primarily driven by the relative abundance of *Prevotella copri*, *Bacteroides vulgatus*, and *Blautia wexerlae*/*Ruminococcus bromii* (Figure S6), which is consistent with the genera that have been linked to enterotypes [37]. However, we found that the samples are distributed continuously along the first three PCs; we did not observe clear clusters in the data. Interestingly, we found that the samples from different host countries are distributed differently along the first three PCs, indicating that individuals from different parts of the world carry different bacterial species in their gut. For example, along PC1 samples from Peru and Tanzania with high levels of *Prevotella copri* clustered together, while along PC2 samples from the United States and China with high levels of *Bacteroides vulgatus* clustered together. This finding is supported by previous work, which found that the relative abundance of *Prevotella copri* could discriminate western and non-western individuals [38]. Despite this taxonomic variation, we were able to able to identify a set of 100 bacterial species with high prevalence (>0.001 relative abundance) across the human population (Figure 3b).

Strain level heterogeneity in the gut is greater between hosts than within hosts

To quantify strain-level heterogeneity within and between individual hosts, we used *PhyloCNV* to identify single-nucleotide variants found in the core genomes of these 100 prevalent gut species across individuals (Methods). Both types of population heterogeneity have been reported in the literature for human gut communities [11-14, 16], but the relative magnitude of each is unknown to our knowledge. To address this question, we estimated nucleotide diversity (π) in the core genome (i.e. genomic regions present in all strains) of each species both within and between individuals, where π is defined as the average pairwise nucleotide difference between two genomes sampled from a population of cells (Methods). Based on this analysis, we found that on average there is 8.5x (range=0.9x to 28.4x) more nucleotide diversity for a species across different individuals than within an individual (Figure 4a). In other words, strains sampled from two different individuals tend to be more diverged than two strains sampled from the same individual. This general trend was common across all the species we examined in each of the different host populations and did not appear to be affected by low sequencing depth in certain samples (Figure S7).

Despite this overall trend, we found that intra-sample strain-level heterogeneity varied by 25x across gut species (range=$4x10^{-4}$ to $1.1x10^{-2}$). For example, *Ruminococcus bromii* had consistently low intra-sample diversity (median π=$6.4x10^{-4}$), whereas *Prevotella copri* and *Faecalibacterium prauznizii* had very high levels of diversity (median π>$4.5x10^{-3}$), indicating the presence of multiple strains per host (Figure 4b). Interestingly, we found only a weak correlation between intra-sample and inter-sample π (R^2=0.14, p=$1x10^{-4}$), as well as a number of species, like *Ruminococcus bromii*, with high diversity across the human population but low diversity within individuals. One possible explanation for these results is competitive exclusion between diverse strains during colonization, which has been observed *in vivo* for *Bacteroides fragilis* [39]. Supporting this hypothesis, we found extremely low intra-sample diversity for *Bacteroides fragilis* (median π=$6.2x10^{-4}$) and several other *Bacteroides* species.

Next, we asked whether intra-sample strain-level heterogeneity varied systematically between host countries. Previous work has shown elevated species-level diversity in individuals living rural lifestyles in South America and Africa [35, 36, 38], but it is not clear whether these findings extend to the strain-level. To address this, we compared the distributions of intra-sample π for each species between host countries and found significant differences for 39% of species tested (Kruskal-Wallis q-value < 0.01). In particular, there was a trend towards increased intra-sample π in individuals from Peru and Tanzania, including *Ruminococcus bromii* and *Facecalbacterium prauznizii* (Figure 4c), indicating that host geography and/or lifestyle may influence strain-level diversity in individuals.

Gene copy number variation is common at short evolutionary timescales in the human gut

We next used *PhyloCNV* to estimate the gene content of strains found in the metagenomes, with the goal of quantifying bacterial genome dynamics, defined as the evolutionary rate of gene gain and loss. Previously, it was reported that strains in the human gut differ in their gene-content by 13% on average [13], but this estimate was made using only 15 reference genomes and did not take into account phylogenetic relatedness of strains between samples.

Based on pan-genome profiling of 100 gut species across our diverse cohort, we found that on average strains from different individuals differ by 20% of their gene content (range=7–47%) and that 15% of genes are gained or lost for every 1% nucleotide divergence in the core genome (range=5–43%) (Methods). However, these are likely underestimates of the true variation since our method cannot detect variation among genes that are present in metagenomes but have not yet been observed in reference genomes. For example, we detected much higher variation in gene content for species with at least 10 sequenced genomes, which differed by 35% between metagenomic samples and had 29% of genes gained or lost for every 1% nucleotide divergence. While this degree of genome plasticity among strains from metagenomes might seem surprising, it is actually less than the variation between reference genomes (Figure S8). Together, these results indicate that members of the same human-associated species can become functionally diverged at relatively short evolutionary timescales and differ significantly in their functional repertoire between hosts.

Bacterial species in the gut differ significantly in their genome dynamics. One of the most dynamic species in metagenomes we analyzed is *Bacteroides ovatus* (Figure 5a), whose gene content differed by 47% on average between metagenomes (range=9–80%) and 51% on average between reference genomes (range=7.5–63%). These differences add up to a total of >3,500 genes that differ in their presence/absence between *B. ovatus* strains, which clearly indicates that different strains must be performing different functions in the gut environment. As expected, we found that strains with less core-genome nucleotide divergence shared more genes than strains with greater divergence (Figure 5b). For example, strains with <0.5% nucleotide divergence shared >68% of genes on average, compared to strains with >2% nucleotide divergence that shared <45% of genes on average.

Next, we classified genes in the *B. ovatus* pan-genome as core, auxiliary, or absent, based on their frequency across metagenomes (>95%, 1-95%, or <1% respectively). We found that core genes (N=1,877) were enriched in basic cellular functions, such as translation (GO:0006412) and nucleotide metabolism (KEGG:00240), and tended to have a characteristic GC content (mean=43.5%, sd=3.4%) (Figures 5c-d). In contrast, auxiliary genes (N=12,445) were enriched in genes of unknown function (FIGfam:00638284), mobile elements (FIGfam:1306568), and genes related to sugar metabolism (KEGG:00520, KEGG:00052, KEGG:00500). Additionally, these genes had a lower GC content (mean=41.7%, t-test p-value < $1\text{x}10^{-16}$) that was more variable across genes (sd=6.8%, Levene's test p-value < $1\text{x}10^{-16}$), and often co-varied together across samples. These features of auxiliary genes are hallmarks of genomic islands [40], and they suggest origins through horizontal gene transfer, which has been previously observed in Bacteroides [41]. The fact that auxiliary genes are enriched in many metabolic functions suggests that the acquisition and loss of these genes might be explained by adaptation to different host environments.

Phylogeography and population structure of prevalent gut species

Finally, we asked whether the strain-level CNV and SNV patterns we observed between samples reflected host geography. Previous studies have found signatures of biogeography for certain members of the human microbiome, such as *Heliobacter pylori* [42] and preliminary evidence for several members of the gut microbiome [11], but a systematic investigation at the global scale is lacking at this time. Systematic differences in the strains carried by different human groups could have implications for using the microbiome in personalized medicine, and more generally would shed light on recent evolutionary history and adaptation of members of the gut microbiome.

To address these questions, we used a standard measure of population differentiation, $F_{ST,}$ to estimate the amount of strain-level genetic variation, based on auxiliary genes or core-genome SNVs, that was explained by host country for each of the prevalent species (Methods). We obtained significant F_{ST} values for 71% of species based on auxiliary genes and for 47% of species based on core-genome SNVs (permutation q-value < 0.01), indicating that most species contained different genes and alleles in different host countries. While we obtained significant F_{ST} values for the majority of species, host geography explained only a minority of the total genetic variation for any of these species (mean F_{ST}=0.15), indicating that many other factors contribute to human-associated bacterial population structure (Figure 6a). To explore which countries contribute the most to the population structure of microbiome species, we computed F_{ST} separately for each pair of host countries and found relatively little differentiation of microbiome bacterial species between countries in Europe and the United States (mean F_{ST}=0.03). In contrast, strains of most bacterial species were differentiated between Tanzania and Peru (mean F_{ST}=0.10) and between these countries and the northern hemisphere countries (mean F_{ST}=0.13).

Next, we compared strain-level biogeography based on CNVs versus SNVs in core genes. Interestingly, we found that CNVs are more correlated with host geography than are SNVs in core genes for the majority of prevalent gut species. This difference may indicate that CNVs can more rapidly change in response to the environment. Alternatively, estimation of CNVs may be more sensitive to technical factors, like read length and sequencing depth, which differ between studies. F_{ST} values based on auxiliary genes and SNVs were highly correlated (R^2=0.91, p-value=$3.9\text{x}10^{-43}$), which provides independent evidence for genomic differentiation between host countries because these data come from separate genomic regions of each species (Figure 6a).

One of the most striking examples of microbiome biogeography is *Eubaterium rectale*, which appears to be structured into four major clades in a phylogeny built from core-genome SNVs (Figure 6b). The different clades occurred with differential frequency across the host countries: clade I was composed of diverged strains from Peruvian and African hosts, clade II was composed of strains from Chinese hosts, and clades III–IV were composed of strains from Americans and Europeans. Interestingly, we were able to recapitulate this population structure by clustering strains based on the fraction of genes shared (Figure 6c). Among genes with the biggest differences in prevalence between countries (F_{ST} > 0.75) were metabolic pathways related to amino sugar and nucleotide sugar metabolism (KEGG:00520), starch and sucrose metabolism (KEGG:00500); enzymes related to degradation of xylan (EC 3.2.1.37, EC 3.2.1.8); and protein families related to polysaccharide transport (FIG00615023, FIG00945504). Together, these results suggest that *Eubaterium rectale* may have adapted to different host environments by the adaptive acquisition of genes to target different available sources of carbohydrates in the human diet. In contrast, we found other species, like *Bacteroides uniformis*, whose population appeared to be structured independently from host geography (Figure S9). We found a number *Bacteroides uniformis* strains from hosts living in different continents that had remarkably few nucleotide differences, indicating that these strains recently shared a common ancestor. More work is needed to better understand the mechanisms underlying differences in biogeography across microbiome species.

Discussion

Recent work has shown extensive strain-level genomic variation of bacteria at the level of gene copy number variants [12, 13] and single nucleotide variants [11], yet there is currently no method to automatically, efficiently, and accurately extract this information from shotgun metagenomes. Existing methods either estimate the relative abundance of known strains [8-10], or use SNVs to phylogenetically type strains [14, 15]. These methods not capture the functions of these organisms and therefore cannot shed light onto the ecological forces shaping their genomes.

To address these issues we developed *PhyloCNV*, which is an integrated computational pipeline that quantifies bacterial species abundance and strain-level genomic variation (i.e. CNVs and SNVs) from shotgun metagenomes. By coupling fast taxonomic profiling via a panel of universal-single-copy genes with sensitive pan-genome and whole genome alignment, *PhyloCNV* can efficiently and automatically compare hundreds of metagenomes to >30,000 genomes clustered into 5,952 species. Our publicly available software and data resources will enable many researchers to conduct strain-level analysis from metagenomes in the future.

Since *PhyloCNV* currently relies on bacterial reference genomes, it cannot quantify variation for novel species, novel genes, or known species from other groups of microbes (e.g. archaea, fungi, and viruses). For these reasons, it will be important to update our database in the future as the number of bacterial reference genomes continues to grow at an exponential pace [43] and to incorporate genomes from other domains of life. Based on the design of our database and algorithm, *PhyloCNV* should scale with this growth of reference data.

We applied our approach to >500 faecal metagenomes from diverse human groups – including the United States, Europe, China, Tanzania, and Peru – and provide the first detailed characterization of global strain-level genomic variation of bacteria in the gut microbiome. We find that current bacterial reference genomes are able to capture a significant proportion of species abundance in the human gut (30-60%), but that most of the strain-level phylogenetic diversity observed in human subjects is novel (90-95%). This result indicates that even in well-studied environments, like the human gut, additional genome sequencing is required in order to establish high quality reference databases.

By leveraging patterns of SNVs found in the core-genome of prevalent species, we find that bacterial strain-level diversity tends to be much greater between individuals than within individuals. A number of species have high levels of diversity in the human population, but low diversity within any individual, which suggests that individual strains of these species colonize their hosts. This is supported by previous work, which found that certain species of *Bacteroides* are resistant to colonization by members of the same species via competitive exclusion [39]. Our results suggest that this process may actually be quite common across gut species and not specific to the *Bacteroides* genus, which could be experimentally validated in the future. If true, this propensity for competitive exclusion has deep implications for microbiome precision medicine. In contrast to this general trend, we found a number of bacterial species with high diversity levels within individuals, indicating that there are likely different mechanisms of colonization across bacterial species. Furthermore, despite low levels of intra-sample diversity across many species, we still observed a large number of low-frequency segregating alleles in these species. It is unclear whether these variants are acquired during the lifespan of the host, as has been shown for certain pathogens [44], or whether they are inherited during colonization. Tracking genomic variation of microbiota between parent and offspring or other interacting individuals will shed light onto selection and demographic processes affecting bacterial populations during colonization.

Using a standard measure of population differentiation, F_{ST}, we found strong evidence for genomic differentiation – based on CNVs and SNVs – between host countries for most of the species examined in this study. In particular, we found that strains from China, Peru, and Africa were more differentiated from strains found in hosts from the Europe and the United States. It is currently unclear the extent to which these patterns are driven by adaptation to differences in the host environment, patterns of host migration, or a combination of factors. Additionally, only a minority of the genetic variation of bacterial species was explained by host geography. It is possible that increased travel between countries has led to mixing of bacterial populations that were at one point highly differentiated. Alternatively other environmental factors (e.g. diet, hygiene, genetics) that vary within human populations might explain the residual intra-species genetic variation. Lastly, these segregating genes and alleles could represent ancient genetic diversity that was present in ancestral bacterial strains and has been maintained across various human populations. As additional metagenomes are sequenced from African hosts, it may be possible to determine the proportion of current diversity that has arisen since migration out of Africa.

Finally, we find that gut communities from Tanzania and Peru are different from other gut communities in several important ways. They have elevated levels of novel species, greater species and strain-level level diversity, and more functionally and phylogenetically diverged strains. Because the hosts in Tanzania and Peru live rural hunter-gather and traditional agricultural lifestyles, their novel species and strains may represent ancestral taxa that have been lost from the microbiomes of humans living industrialized lifestyles. Furthermore, the increased species and strain level diversity might be explained by less hygiene and reduced antibiotics exposure in these rural populations. As additional metagenomes are sequenced from diverse global populations, it will be possible to disentangle the degree to which these patterns are specific to particular lifestyles and/or geography. If hosts living more traditional lifestyles and with more limited access to health care indeed have more diverse microbiomes regardless of host migratory patterns, further exploration of strain-level variation with *PhyloCNV* may shed light on the “hygiene hypothesis” and the role that loss of ancestral microbiome diversity may play in the rise of autoimmune disease in industrialized countries.

While humans have been living with microbes throughout our evolution, we are just beginning to understand global patterns of variation. An understanding of how bacterial strains vary within and between hosts provides a necessary foundation for future studies and for linking the microbiome to human health and disease.

Materials and Methods

Shotgun metagenome simulations

To validate and optimize *PhyloCNV* we designed a series of realistic metagenomic simulations using reads from completed genome-sequencing projects deposited in the NCBI SRA [45] (Table S7). We used this data to construct 20 mock metagenomes, which each contained Illumina reads from 20 randomly selected Bacterial genome projects (Table S8). We trimmed these reads in order to construct metagenomes with read lengths of 100 base pairs. Paired-end reads were sampled from each genome project such that the coverage of each genome was proportional to its relative abundance. We simulated libraries that contained 100x total genome coverage, and the relative abundances of the 20 genomes were distributed with exponentially decreasing relative abundances (50%, 25%, 12%, 6.5% etc.).

Construction of genome-clusters

We followed a protocol similar Mende et al. [18] to cluster bacterial genomes into species groups. We began with 33,252 prokaryotic genomes downloaded from PATRIC (March 2015) [17] and identified one-to-one protein homologs of 112 bacterial marker gene families [21] across the genomes with E-values ≤1e-5 using HMMER3 [46]. Low quality genomes were removed that contained fewer than 100 marker genes (N=1,837) or contained greater than 1,000 contigs (N=618), which left us with 31,007 high-quality genomes. Next, we performed an all-versus-all BLASTN [47] for each set of marker gene homologs and filtered local alignments where either the query or target was covered by <70% of its length. These alignments were used to measure the distance between each pair of genomes: $D_{ab} = (100 - P_{ab})/100$, where P_{ab} was the percent identity of a marker gene between genomes a and b. Next, we performed average-linkage clustering using the program MC-UPGMA [48] and applied a range of distance thresholds to identify genome-clusters.

For validation, we compared our results to average nucleotide identity (ANI), which is considered to be a gold standard for delineating prokaryotic species [22, 23] but was too computationally intensive to compute for all genome-pairs. Specifically, we used the procedure described by Richter and Rossello-Mora [22] to compute ANI for >18,000 genome-pairs (Table S6) and labeled pairs of genomes with ANI ≥ 95% as members of the same species and pairs of genomes with ANI < 95% as members of different species. We compared these labels to our genome-clusters and classified each genome-pair into one of the following categories: *True positive*: a clustered genome-pair with ANI ≥ 95%; *False positive*: a clustered genome-pair with ANI < 95%; *False negative*: a split genome-pair with ANI ≥ 95%; *True negative*: a split genome-pair with ANI < 95%. Based on these classifications we calculated the true positive rate (TPR), precision (PPV), and F1-score for each clustering result (Table S2). Based on this evaluation, we identified a subset of 30 gene families that had good agreement with ANI (all F1 > 0.93), and to increase clustering performance, we combined pairwise distances across these gene-families and re-clustered genomes using MC-UPGMA (Table S3). We found that a distance cutoff of 0.035 (96.5% identity) maximized the F1-score at 0.98 and resulted in 5,952 genome-clusters (Table S3). Finally, we annotated each genome-cluster by the most common Latin name within the cluster (Table S4).

Genomic database construction

Genome-clusters were leveraged in order to compile a comprehensive genomic data resource used by *PhyloCNV*. First, we identified a single representative genome from each genome-cluster to use for detecting single-nucleotide variants (5,952 genomes). Each representative genome was chosen to minimize its phylogenetic distance to other members of the genome-cluster. Next, we extracted a panel of 15 universal single-copy genes from each representative genome (88K genes) to use for rapidly estimating the abundance of genome-clusters from shotgun metagenomes. Marker genes were selected based on their ability to accurately recruit metagenomic reads as well being universal and single-copy. Finally, we used USEARCH [49] to identify the set of unique genes at 99% identity across the genomes within each genome-cluster to use for metagenomic pan-genome profiling. Overall, this procedure resulted in clustering of 116,978,184 genes from 32,007 genomes into 31,840,245 non-redundant genes. We further clustered these non-redundant genes at different levels of sequence identity (75-95% identity) in order to identity *de novo* gene families of varying size and diversity for downstream analyses. Functional annotations for pan-genomes were obtained from PATRIC (March 2015) [17] and include: FIGfams [24], Gene Ontology [25], and KEGG Pathways [26].

Estimation of species abundance from shotgun metagenomes

Next, we developed a method to estimate the abundance of genome-clusters from shotgun metagenomes. First, we performed a simulation experiment to identify the subset of marker genes that were able to accurately recruit reads to the correct genome-clusters. We simulated one hundred 100-bp reads from each of the 112 marker genes in each of the 5,952 genome-clusters and used HS-BLASTN [50] to map these reads back to a database that contained the full length marker gene sequences. To simulate the presence of novel species and strains, we discarded alignments between reads and reference sequences from the same genome-cluster, assigned each read to the genome-cluster according to it's top-hit, and measured recruitment performance using the F1-score. Based on this experiment, we identified 15 universal single-copy marker genes that were able to accurately recruit metagenomic reads (Table S9). Additionally, we identified optimal gene-specific species-level % identity cutoffs for mapping reads to the databases, which ranged from 94.5% to 98.00% identity (Table S9).

To perform taxonomic profiling, *PhyloCNV* aligns reads to the database of 15 marker genes with HS-BLASTN, discards local alignments that cover <70% of the read or alignments that fail to satisfy the gene-specific species-level % identity cutoffs, and assigns each uniquely-mapped read to the genome-cluster according to it's best-hit. *PhyloCNV* then probabilistically assigns non-uniquely mapped reads (i.e. identical alignment scores to marker genes from >1 genome-cluster) based on the relative abundances of genome-clusters estimated from uniquely mapped reads. These mapped reads are then used to estimate the coverage of each genome-cluster: $C_j = \frac{N_k \times L_k}{L_j}$, where N_k is the number of mapped reads, L_k is the alignment length of mapped reads, and L_j is the total gene length across the 15 marker genes from the genome-cluster. To optionally estimate cellular relative abundances, R_j, *PhyloCNV* divides the genome-coverage of each genome-cluster by the total coverage across all microbial genomes present in the metagenome, C_k, which is estimated using *MicrobeCensus* [27]: $R_j = C_j / C_k$. The fraction of novel species abundance in a community is estimated by taking the sum of coverages over known species, and dividing this total by the total coverage across all microbial genomes: $\sum_j^n C_j / C_k$.

PhyloCNV was validated on 20 simulated 100x metagenomes that were each composed of 100-bp reads from 20 distinct species with staggered relative abundances (Tables S7-8). The speed of *PhyloCNV* was compared to MetaPhlAn and mOTU using default parameters for all methods using 1 million reads from 12 different metagenomes from different environments, including marine [ERR594313, ERR594318, ERR594337], human gut [SRR413720, SRR413711, SRR413707], primate gut [SRR1747029, SRR1747061, SRR1747021], and soil [SRR350969, SRR400520, SRR400521].

Estimation of pan-genome profiles from shotgun metagenomes

Next, we developed a framework to estimate the pan-genomic content of abundant microbial species from shotgun metagenomes. In the first step of the pipeline, *PhyloCNV* dynamically builds a pan-genome database that contains all 99% non-redundant genes from the subset of species that are present in the metagenome. By integrating taxonomic information, our method avoids aligning reads to organisms that are either absent or undetectable. Next, our framework uses Bowtie2 [51] to locally map reads from the input metagenome against the pan-genome database. Each read is mapped a single time according to its best hit, and alignments with an insufficient number of matched nucleotides or insufficient alignment coverage are discarded. Mapped reads are then used to compute the coverage of each 99% non-redundant gene across the pan-genomes: $C_g = \frac{N_k \times L_k}{L_g}$, where N_k is the number of mapped reads, L_k is the read length, and L_g is the gene length. These gene coverages are optionally aggregated to obtain coverages of larger gene-families at lower sequence identity (75-95%). This feature gives users the flexibility to estimate the coverage of gene families of varying size and diversity, while maintaining mapping speed and sensitivity. Finally, gene coverages are normalized by the median coverage across a panel of 15 universal single-copy genes, resulting an estimated copy-number of the gene in a given microbial species in a given sample.

To evaluate performance, we ran our method on 20 simulated 100x metagenomes that were each composed of 100-bp reads from 20 distinct species with staggered relative abundances (Tables S7-8). We used default settings and estimated the copy-number of gene families clustered at 90% identity. To predict gene presence/absence, we applied a threshold to gene copy number values and classified genes with values above the threshold as present and those below the threshold as absent. True positives were present genes predicted as present, false positives were absent genes predicted as present, true negatives were absent genes predicted as absent, and false negatives were present genes predicted as absent. Performance was evaluated using the statistic accuracy, which was obtained by counting number of correct classifications (TP+TN) divided by the total number of genes (TP+TN+FN+FP). We found that a gene copy number cutoff of ~0.35 maximizes accuracy at 0.98 across the 20 simulations.

Identification of single-nucleotide variants

Finally, we developed a framework to identify single-nucleotide variants (SNVs) within the genomes of abundant microbial species from shotgun metagenomes. As before, our framework first constructs a database that contains representative genomes from the subset of species that are present in the metagenome and reads are globally aligned against the reference genome database. Each read is mapped a single time according to its best hit, and alignments with an insufficient number of matched nucleotides are discarded. Next, samtools [52] is used to filter low quality bases according to MAPQ and BASEQ scores, generate pileups, and call single-nucleotide variants.

To evaluate performance, we ran our method on 20 simulated metagenomes composed of Illumina reads from bacterial genome sequencing projects and called consensus SNVs within strains in our simulated metagenomes (Tables S7-8). We focused on a subset of 71 "novel" strains that were distinct from the reference genomes used for read mapping. To identify true SNVs, we used the genome-alignment program MUMmer [53], which identified 3,971,528 true substitutions across the 71 strains (mean=55,937). Next, we compared the predicted SNVs generated by our pipeline to the true SNVs identified by MUMmer and evaluated performance using the true positive rate (i.e. fraction of total SNVs that were correctly predicted) and precision (i.e. fraction of predicted SNVs that were correct). Additionally, we estimated the ANI of novel strains to sequenced representatives and compared these estimates to expected values generated by MUMmer.

Application of PhyloCNV to public datasets

We applied *PhyloCNV* to 544 stool metagenomes from healthy unrelated subjects from seven different studies, totaling over 29 billion reads and 2.6 tera base pairs. This data included individuals from Austria [ERP008729], Tanzania [SRP056480], Peru [SRP052307], Sweden [ERP002469], China [SRP011011], Spain and Denmark [ERP004605], and the United States [SRP002163]. All datasets were identified using the SRAdb relational database [54]. First, we used *PhyloCNV* to taxonomically profile each sample, using up to 10 million single-end reads from each sample. Based on these results we identified the 100 most prevalent species in the gut and performed pan-genome profiling and SNV prediction on these species across the 544 samples. To reduce false-positives, we discarded alignments with <94% identity or alignments covering <75% of a read's length. Additionally, for SNV prediction, we discarded bases with a map quality score of <20 or a base quality score of <25. To further improve data quality, we discarded low-coverage samples with <5x depth at marker-genes or samples where the reference genome was covered by <40% of its length. Next, we computed coverage of genes clustered at 90% identity and classified genes with an estimated copy number ≥0.35 as present and those below as absent, which was identified as the optimal cutoff based on our simulations. Additionally, for SNVs, we called consensus alleles at each site covered by at least 3 reads. Finally, we merged the output from *PhyloCNV* across samples for each of the 100 genome-clusters.

Identification of core genes and genomic regions

For each species we used CNVs to classify pan-genome genes as either core, auxiliary, or absent based on their prevalence across samples. We defined core genes as occurring in ≥95% of samples, auxiliary genes as occurring in 1–95% of samples, and absent genes as occurring in <1% of samples. Separately, we used SNVs to identify core genome sites in each representative genome. A core genomic site was defined as a position in the representative genome with ≥3 mapped reads in ≥95% of samples.

Constructing strain-level phylogenies

We used *PhyloCNV* to build strain-level phylogenies for each of the 100 gut species. For each species, we called consensus alleles at each core-genome site for each sample. Additionally, we called alleles at these sites for all available reference genomes. This resulted in a FASTA file for each species containing one consensus sequence for each sample and each reference genome. Next, we used FastTree [55] to build approximately maximum-likelihood trees for each of the species, where each tip in the tree represents a particular strain found in a particular sample or a reference genome. No multiple sequence alignment was necessary, because SNVs are all annotated relative to the same reference sequence.

For each strain-level phylogenic tree, we estimated the amount of phylogenetic diversity [56] shared between reference genome(s) and strains found in metagenomic samples using the following equation: $PD_{shared} = PD_{ref} + PD_{meta} - PD_{ref+meta}$, where PD_{ref} and PD_{meta} are sub-trees that contain either reference genomes or metagenomes, and $PD_{ref+meta}$ is the phylogenetic tree containing metagenomes and reference genomes. We then estimated the fraction of phylogenetic diversity found in metagenomic samples that was shared with reference genomes: $F = PD_{shared}/PD_{meta}$.

Estimation of intra and inter sample nucleotide diversity

We used SNV frequencies to estimate the core-genome nucleotide diversity of each species, π, using the following equation: $\pi = \frac{1}{n}\sum_i^n 2p_i q_i$, where p_i is the reference allele frequency at core-genome site *i*, q_i=1 $-$ p_i, and *n* is the length in base pairs of the core-genome. This equation is adapted from [57] where it was used to estimate nucleotide diversity from human polymorphism data. Here, π is defined as the average number of nucleotide differences between two genomes sampled from a population of cells. To estimate the intra-sample nucleotide diversity, π_{within}, we used reference allele frequencies estimated from individual metagenomic samples in the above equation. To summarize the estimates for a given species across samples, we took the median. To estimate the inter-sample nucleotide diversity, $\pi_{between}$, we called consensus alleles at each core-genome site for each sample, estimated allele frequencies across samples, and used these allele frequencies to estimate π.

Estimation of gene-content difference, nucleotide divergence, and genome dynamics

We used the Jaccard coefficient to estimate the proportion of genes that differed between each pair of metagenomes and each pair of reference genomes for each bacterial species. Additionally, we used strain-level phylogenetic trees to estimate patristic distances (i.e. nucleotide divergence) between metagenomes and between reference genomes for each species. We used the ratio of these estimates (gene-content difference/nucleotide divergence) to estimate the percentage of genes that were gained or lost between strains for every 1% nucleotide divergence in the core-genome.

Estimation of F_{ST} for CNVs and SNVs

We estimated the fixation index (F_{ST}) – defined as the amount of genetic variation explained by population structure – using both CNVs and SNVs. For SNVs, we computed F_{ST} using inter-sample allele frequencies at core-genome sites, and for CNVs, we computed F_{ST} using inter-sample gene presence/absence frequencies: $F_{ST} = \frac{\bar{p}(1-\bar{p}) - \sum c_i p_i (1-p_i)}{\bar{p}(1-\bar{p})}$, where p_i is the allele or gene frequency in subpopulation *i*, c_i is the relative size of subpopulation *i*, and $\bar{p}$ is the average allele or gene frequency in the total population. In our analysis, each subpopulation was a different host country (United States, China, Austria, Denmark, Spain, Sweden, Peru, and Tanzania). To obtain an F_{ST} value for an individual species, we took a diversity weighted average of F_{ST} values across alleles or genes, and to obtain a p-value for species F_{ST} values, we performed a permutation test, in which the host country labels were shuffled.

Figure legends

Figure 1. A comprehensive data resource for comparative microbial (meta)genomics
A) 31,007 genomes were hierarchically clustered based on the pairwise identity across a panel of 30 phylogenetic marker genes (pMGs). 5,952 species groups were identified by applying a 96.5% identity cutoff. **B)** Comparison of genome-clusters to annotated species names. Out of 31,007 genomes assigned to a genome-cluster, 5,701 (18%) disagreed with the taxonomic label. Most disagreements are due to genomes lacking annotation at the species level (47%). **C)** Genome-clusters were leveraged to construct three genomic databases to be used for species and strain-level profiling of microbial communities. Arrows denote genes with colors indicating gene families. Non-redundant Pan Genomes: the set of unique (>=99% identity) genes from each genome-cluster. Representative genomes: the most phylogenetically representative genome from each genome-cluster. Phylogenetic marker genes: a set of 15 universal single-copy marker genes from each genome-cluster, which are capable of accurately recruiting metagenomic reads.

Figure 2. An integrated pipeline for profiling species abundance and strain-level genomic variation from metagenomes
A) Metagenome species profiling. Reads from a metagenomic sample are aligned against a database of pMGs and are assigned to species groups according to gene-specific and species-level cutoffs. Mapped reads are used to estimate the genome-coverage and relative abundance of 5,952 genome-clusters. **B)** Metagenome pan genome profiling. In the second step of the pipeline, a pan genome database is dynamically constructed based on the subset of genome-clusters that are present at high coverage (e.g. >1x) in the metagenome. Metagenomic reads are locally aligned to the gene database using Bowtie2 and mapped reads are used to obtain pan gene coverages, which are normalized by the median coverage across a panel of 15 universal single-copy genes. **C)** Metagenome single-nucleotide variant prediction. In the third step of the pipeline, a representative genome database is constructed, as described in (B). Metagenomic reads are globally aligned to the genome database using Bowtie2 and mapped reads are used to identify variants, predict consensus alleles, and estimate allele frequencies. **D)** For each genome-cluster, results are merged across one or more samples to generate several outputs. For pan genome analysis, the outputs include a gene presence/absence matrix and a gene copy number matrix. For SNV analysis, the outputs include an allele frequency matrix, core-genome consensus sequences, and an approximate maximum-likelihood phylogenetic tree, which optionally includes phylogenetic placements of sequenced reference genomes.

Figure 3. Identification of prevalent gut species that span multiple host populations. *PhyloCNV* was applied to 539 stool metagenomes from healthy unrelated subjects from from the United States, China, Europe (Austria, Denmark, Spain, Sweden), Tanzania, and Peru. **A)** The % of novel species cellular relative abundance was estimated for each sample. Plotted is the distribution of this statistic across host countries. A high value indicates the abundance of species that were not captured by the reference database. **B)** Based on metagenomic species profiling, we estimated the prevalence of genome-clusters across samples. Prevalence is defined as the presence of a genome-cluster in a sample with a cellular relative abundance of >1/1000. For display purposes, only the top 30 genome-clusters are shown.

Figure 4. Strain level heterogeneity is greater between hosts than within hosts.
A) We estimated the intra-sample and inter-sample nucleotide diversity for each prevalent species. Distributions of intra-sample diversity are shown in boxplots. The inter-sample diversity of each species is indicated in red. On average there is 8.5x more strain-level diversity between samples than within samples. For display purposes, only the top 25 most prevalent species are shown. **B)** Minor allele frequency (MAF) distributions for two selected species: *Faecalibacterium prausnitzii* and *Ruminococcus bromii*. Nucleotide diversity is indicated on the right. Samples 1-15 indicate intra-sample MAF distributions for 15 randomly selected samples for each species. 'Between samples' indicates minor allele frequencies estimated across samples. **C)** Distributions of intra-sample nucleotide diversity for *Faecalibacterium prausnitzii* and *Ruminococcus bromii* across host populations.

Figure 5. Extensive differences in the gene content of *Bacteroides ovatus* strains between hosts.
A) Presence or absence of genes in the *Bacteroides ovatus* pangenome across human faecal metagenomes. Column colors indicate whether a gene is core (blue), auxiliary (red), or absent (green). **B)** Gene content difference between pairs of metagenomes as a function of phylogenetic distance. More genes are shared between strains with less nucleotide divergence in the core genome. **C)** Gene set enrichment analysis was performed to identify overrepresented functional categories in the core genome, auxiliary genome, and genes that only occur in reference genomes. **D)** Gene nucleotide composition of different classes of genes in the *Bacteroides ovatus* pangenome. Auxilliary and absent genes have a different distribution of GC content.

Figure 6. Phylogeography of *Eubacterium rectale* in the human gut.
A) We computed F_{ST} for each species based on core genome SNVs or CNVs (i.e. auxiliary genes). Estimates of F_{ST} based on CNV and SNV frequency are highly concordant. **B)** Core-genome phylogenetic tree for *Eubacterium rectale* strains from metagenomes and reference genomes. Colored tips indicate phylogenetic placements of strains found in metagenomic samples. Labeled tips indicate phylogenetic placements of reference genomes. **C)** Gene sharing between *E. rectale* strains found in metagenomic samples. Row and columns colors indicate phylogenetic clades defined in (B). The population structure of *E. rectale* based on the core-genome and auxiliary genome is consistent.

Acknowledgements

We would like to thank Nandita Garud for reading the manuscript and providing helpful feedback. This project was supported by funding from NSF grant #DMS-1069303, Gordon & Betty Moore Foundation grant #3300, the San Simeon Fund, and institutional funds from Gladstone Institutes. Author contributions

Author contributions

SN designed and implemented the algorithm, SN and KSP designed the experiments, and SN performed the analysis. SN and KSP wrote the paper. Both authors read and approved the final manuscript.

References

1. Henry J. Haiser DBG, Kelly Chatman, Gopal Sirasani, Emily P. Balskus, Peter J. Turnbaugh: **Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium Eggerthella lenta.** *Science* 2013, **341**.
2. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G: **The shared antibiotic resistome of soil bacteria and human pathogens.** *Science* 2012, **337:**1107-1111.
3. Evan S. Snitkin AMZ, Clemente I. Montero, Frida Stock, Lilia Mijares, NISC Comparative Sequence Program, Patrick R. Murray, and Julie A. Segre: **Genome-wide recombination drives diversification of epidemic strains of Acinetobacter baumannii.** *PNAS* 2011, **108**.
4. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C: **The Salmonella enterica pan-genome.** *Microb Ecol* 2011, **62:**487-504.
5. LeBlanc JJ: **Implication of virulence factors in Escherichia coil O157:H7 pathogenesis.** *Crit Rev Microbiol* 2003, **29:**277-296.
6. Janda JM, Abbott SL: **16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.** *J Clin Microbiol* 2007, **45:**2761-2764.
7. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al: **Metagenomic species profiling using universal phylogenetic marker genes.** *Nat Methods* 2013, **10:**1196-1199.
8. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje GB, Clement MJ, Crandall KA, Johnson WE: **Pathoscope: species identification and strain attribution with unassembled sequencing data.** *Genome Res* 2013, **23:**1721-1729.
9. Ahn TH, Chai J, Pan C: **Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance.** *Bioinformatics* 2015, **31:**170-177.
10. Tu Q, He Z, Zhou J: **Strain/species identification in metagenomes using genome-specific markers.** *Nucleic Acids Res* 2014, **42:**e67.
11. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, et al: **Genomic variation landscape of the human gut microbiome.** *Nature* 2013, **493:**45-50.
12. Greenblum S, Carr R, Borenstein E: **Extensive strain-level copy-number variation across human gut microbiome species.** *Cell* 2015, **160:**583-594.
13. Zhu A, Sunagawa S, Mende DR, Bork P: **Inter-individual differences in the gene content of human gut bacterial species.** *Genome Biol* 2015, **16:**82.
14. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D: **ConStrains identifies microbial strains in metagenomic datasets.** *Nat Biotechnol* 2015.
15. Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P: **Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data.** *Genome Med* 2015, **7:**52.
16. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF: **Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization.** *Genome Res* 2013, **23:**111-120.

17. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, et al: **PATRIC, the bacterial bioinformatics database and analysis resource.** *Nucleic Acids Res* 2014, **42:**D581-591.

18. Mende DR, Sunagawa S, Zeller G, Bork P: **Accurate and universal delineation of prokaryotic species.** *Nat Methods* 2013, **10:**881-884.

19. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A: **Microbial species delineation using whole genome sequences.** *Nucleic Acids Res* 2015.

20. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L: **Update on RefSeq microbial genomes resources.** *Nucleic Acids Res* 2015, **43:**D599-605.

21. Wu D, Jospin G, Eisen JA: **Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups.** *PLoS One* 2013, **8:**e77033.

22. Richter M, Rossello-Mora R: **Shifting the genomic gold standard for the prokaryotic species definition.** *Proc Natl Acad Sci U S A* 2009, **106:**19126-19131.

23. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**1929-1940.

24. Meyer F, Overbeek R, Rodriguez A: **FIGfams: yet another set of protein families.** *Nucleic Acids Res* 2009, **37:**6643-6654.

25. Consortium TGO: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25.**

26. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28.**

27. Nayfach S, Pollard KS: **Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome.** *Genome Biol* 2015, **16:**51.

28. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes.** *Nat Methods* 2012, **9:**811-814.

29. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al: **An integrated catalog of reference genes in the human gut microbiome.** *Nat Biotechnol* 2014, **32:**834-841.

30. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J, Backhed F: **Gut metagenome in European women with normal, impaired and diabetic glucose control.** *Nature* 2013, **498:**99-103.

31. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, et al: **Gut microbiome development along the colorectal adenoma-carcinoma sequence.** *Nat Commun* 2015, **6:**6528.

32. Consortium THMP: **A framework for human microbiome research.** *Nature* 2012, **486:**215-221.

33. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al: **A metagenome-wide association study of gut microbiota in type 2 diabetes.** *Nature* 2012, **490:**55-60.

34. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, Brigidi P, Crittenden AN, Henry AG, Candela M: **Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.** *Curr Biol* 2015, **25:**1682-1693.

35. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, et al: **Subsistence strategies in traditional societies distinguish gut microbiomes.** *Nat Commun* 2015, **6:**6505.

36. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, et al: **Gut microbiome of the Hadza hunter-gatherers.** *Nat Commun* 2014, **5:**3654.

37. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al: **Enterotypes of the human gut microbiome.** *Nature* 2011, **473:**174-180.

38. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al: **Human gut microbiome viewed across age and geography.** *Nature* 2012, **486:**222-227.

39. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK: **Bacterial colonization factors control specificity and stability of the gut microbiota.** *Nature* 2013, **501:**426-429.

40. Langille MG, Hsiao WW, Brinkman FS: **Detecting genomic islands using bioinformatics approaches.** *Nat Rev Microbiol* 2010, **8:**373-382.

41. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA: **Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon.** *Appl Environ Microbiol* 2001, **67:**561-568.

42. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al: **Traces of human migrations in Helicobacter pylori populations.** *Science* 2003, **299:**1582-1585.

43. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, et al: **Insights from 20 years of bacterial genome sequencing.** *Funct Integr Genomics* 2015, **15:**141-161.

44. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R: **Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures.** *Nat Genet* 2014, **46:**82-87.

45. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: **The sequence read archive.** *Nucleic Acids Res* 2011, **39:**D19-21.

46. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol* 2011, **7:**e1002195.

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.

48. Loewenstein Y, Portugaly E, Fromer M, Linial M: **Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space.** *Bioinformatics* 2008, **24:**i41-49.

49. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26:**2460-2461.

50. Chen Y, Ye W, Zhang Y, Xu Y: **High speed BLASTN: an accelerated MegaBLAST search tool.** *Nucleic Acids Res* 2015.

51. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

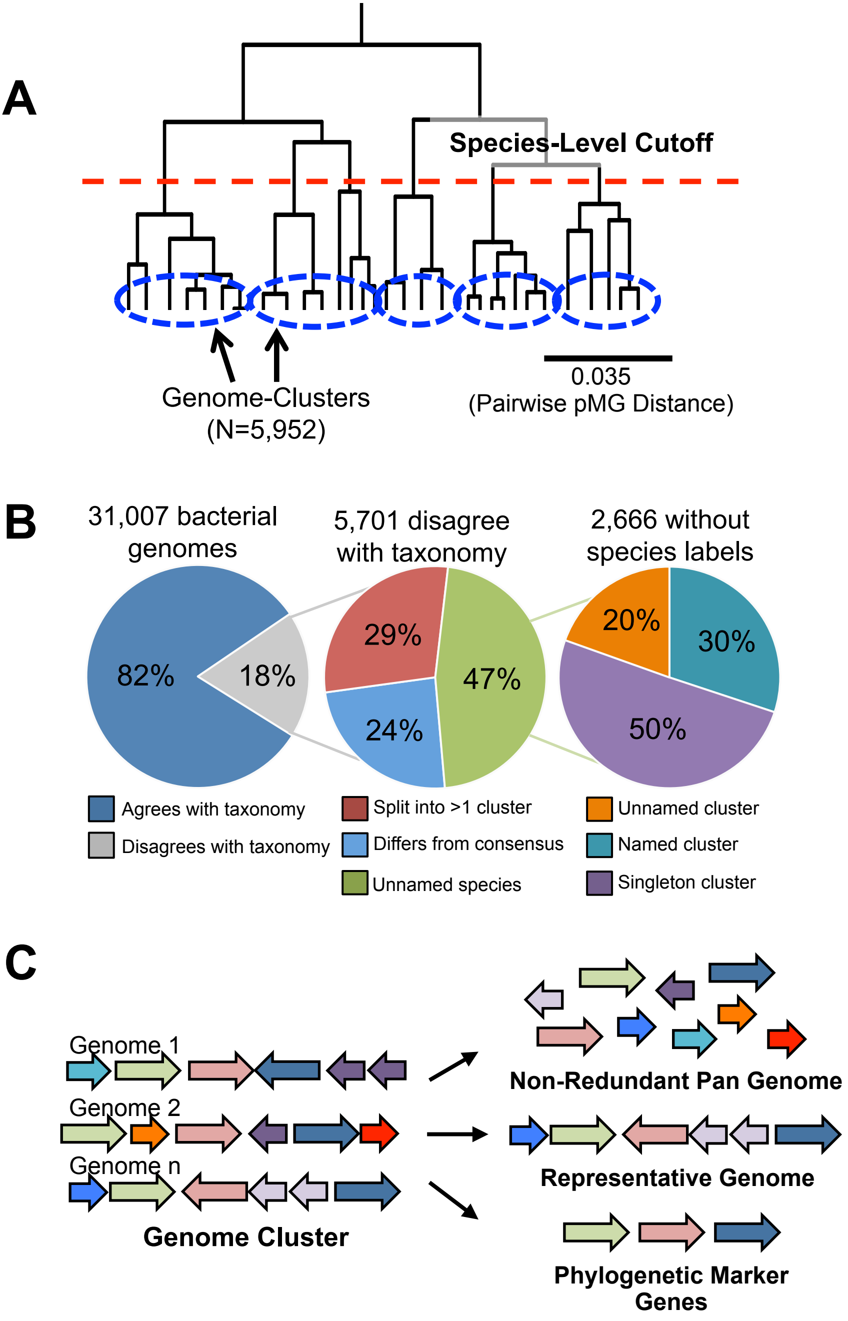
53. Stefan Kurtz AP, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg: **Versatile and open software for comparing large genomes.** *Genome Biology* 2004, **5**.

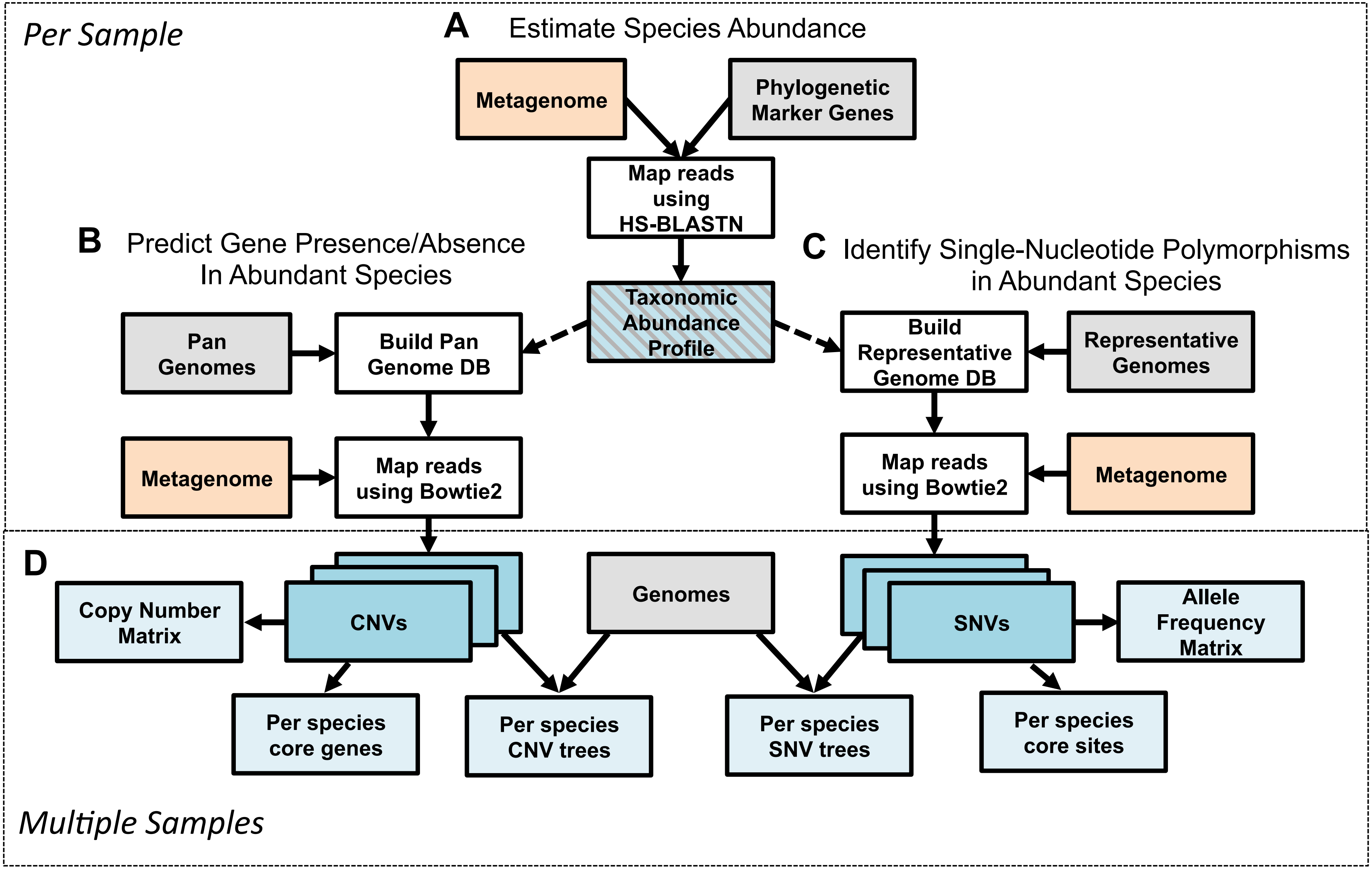
54. Yuelin Zhu RMS, Paul S Meltzer, Sean R Davis: **SRAdb: query and use public next-generation sequencing data from within R.** *BMC bioinformatics* 2013, **14:**19.

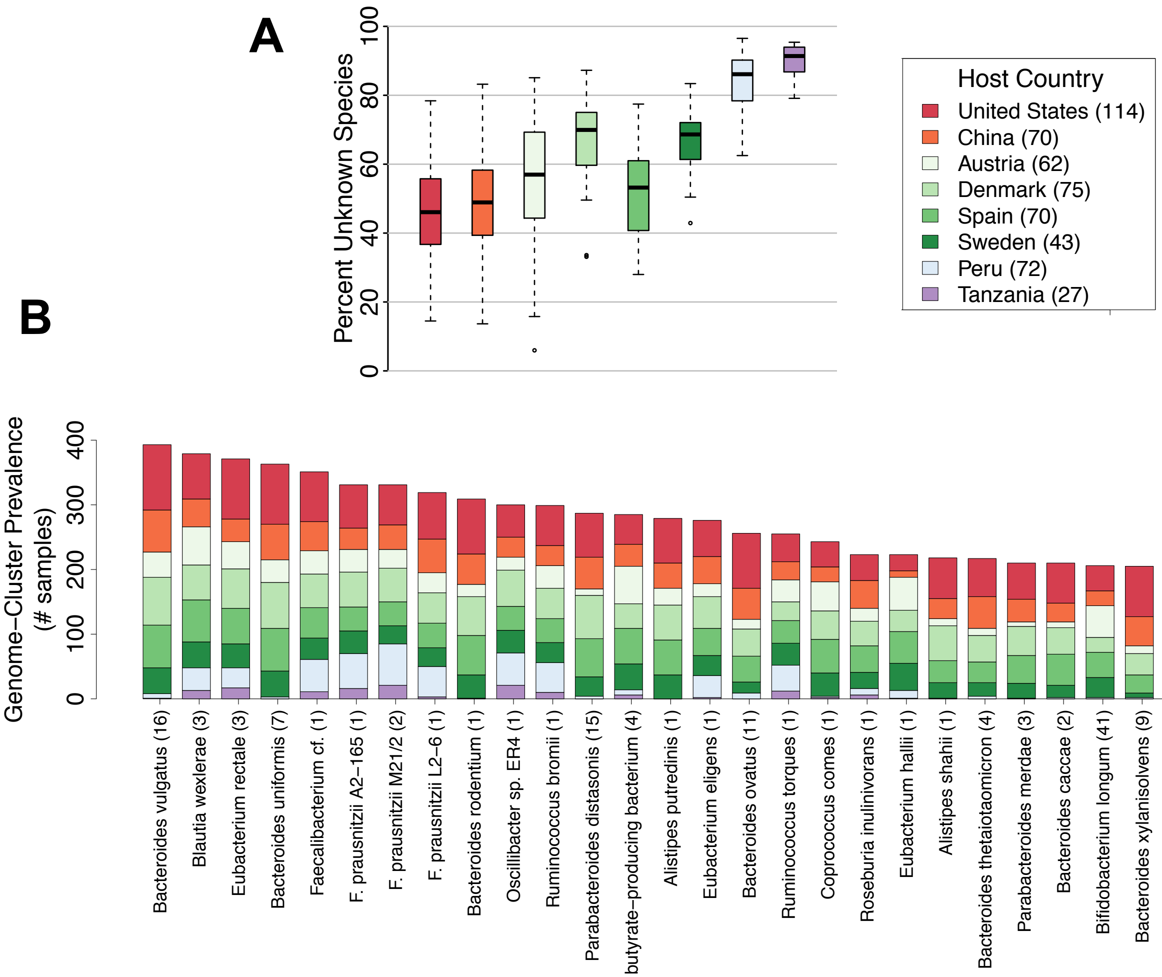
55. Morgan N. Price PSD, Adam P. Arkin: **FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.** *PLoS ONE* 2010, **5**.

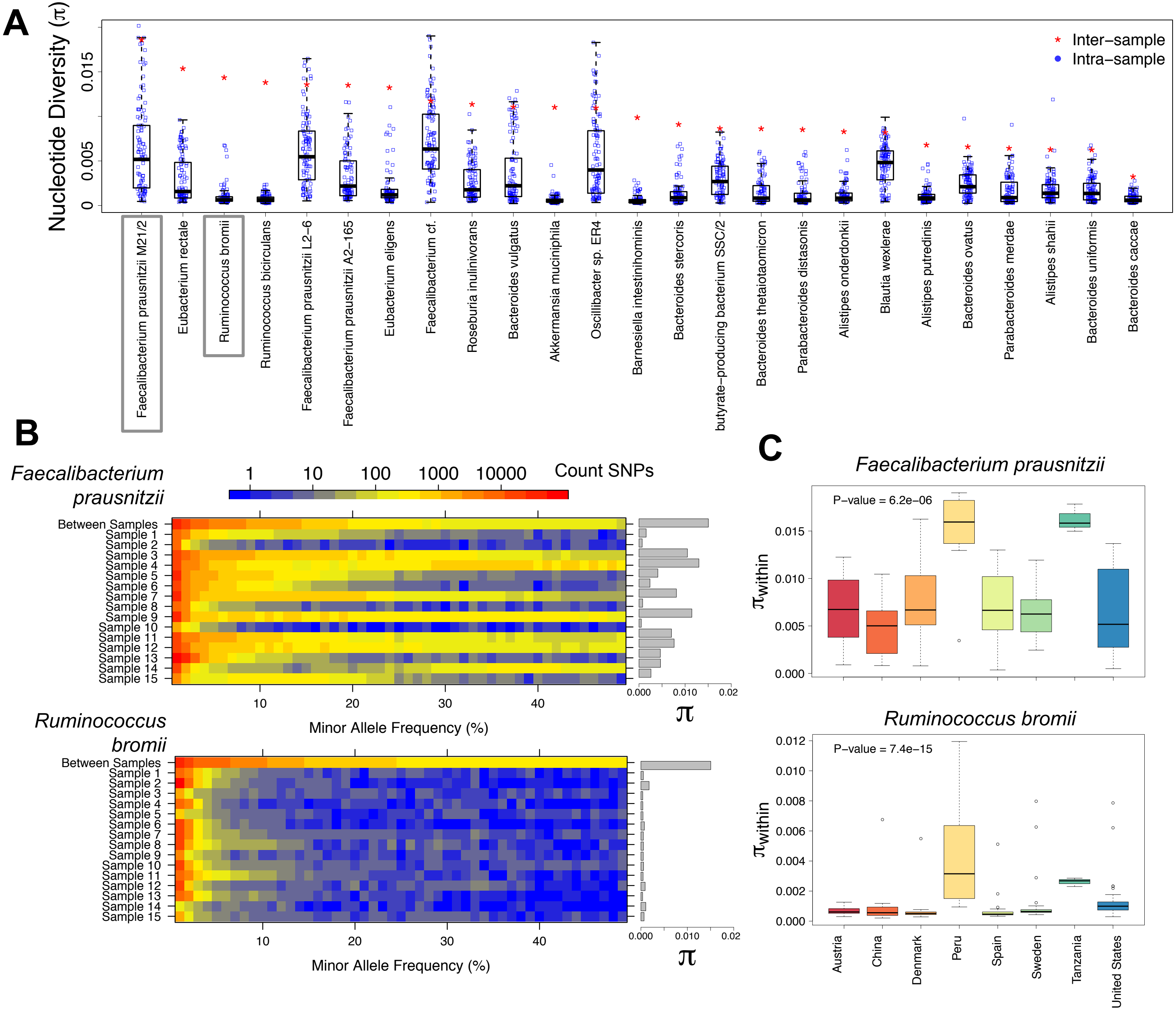
56. Faith DP: **Conservation evaluation and phylogenetic diversity** *Biological Conservation* 1991, **61.**

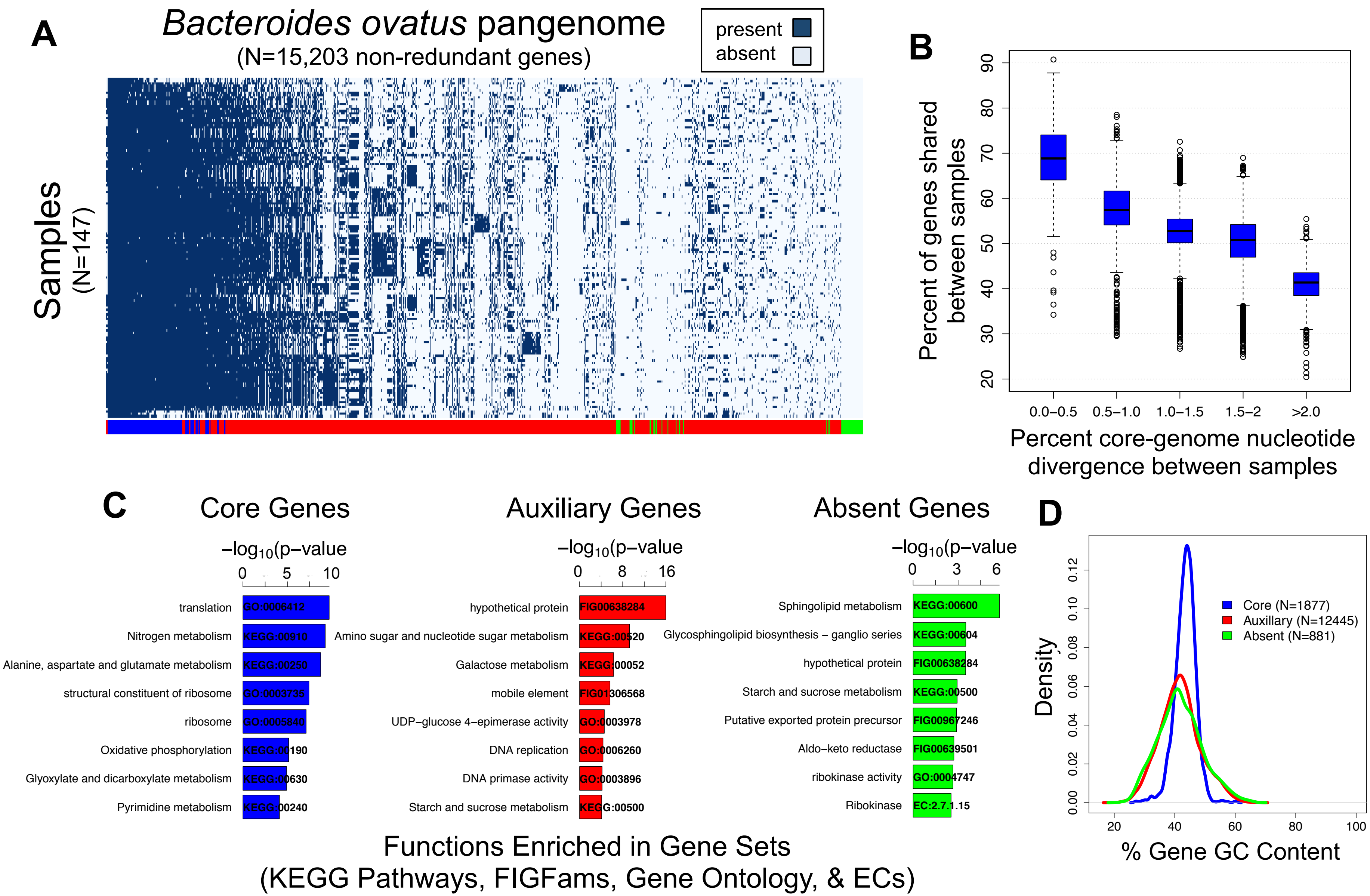
57. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB: **Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project.** *Nucleic Acids Res* 2011, **39:**7058-7076.

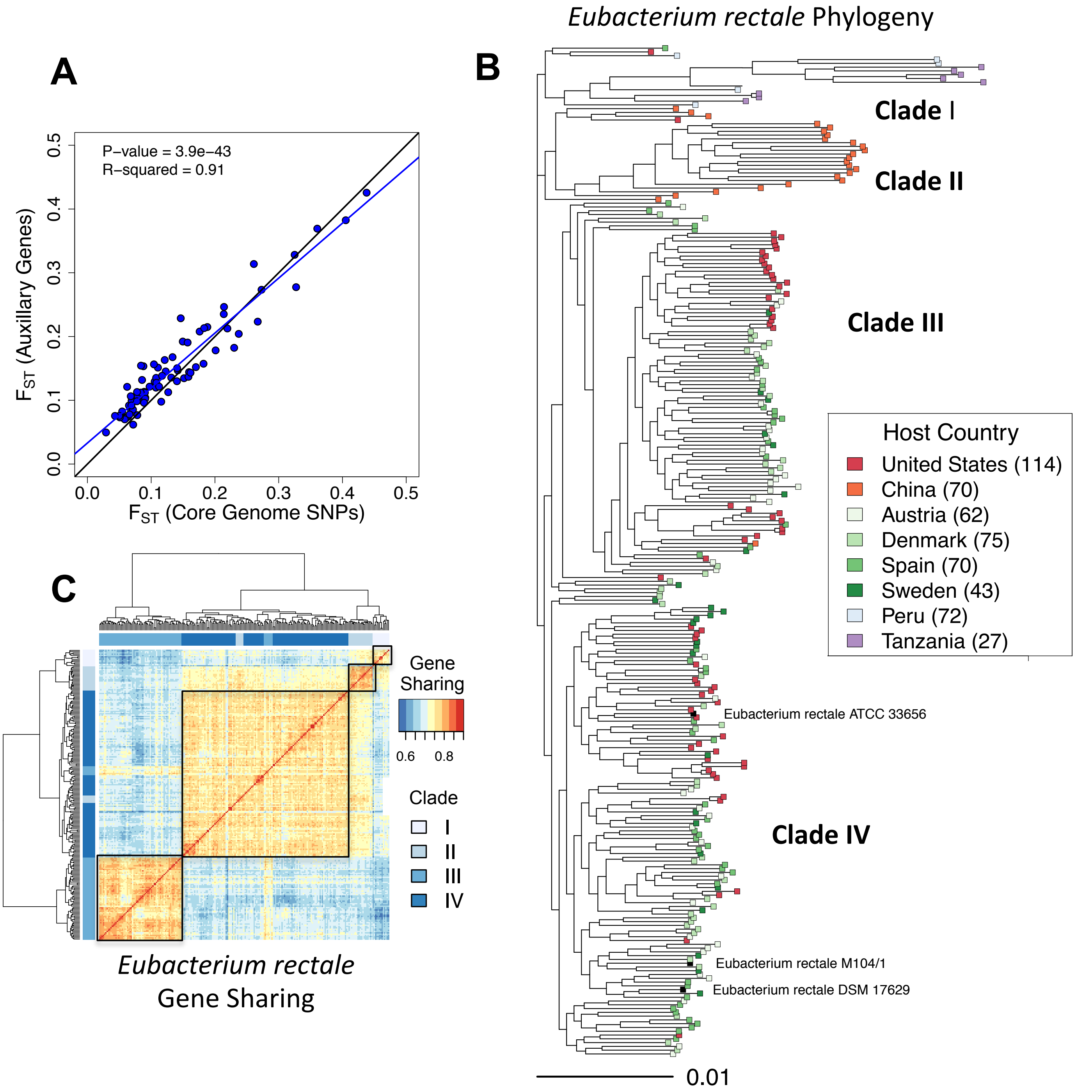

A
Species-Level Cutoff
Genome-Clusters
(N=5,952)
0.035
(Pairwise pMG Distance)
B
31,007 bacterial genomes
5,701 disagree with taxonomy
2,666 without species labels
82%
18%
29%
24%
47%
20%
30%
50%
Agrees with taxonomy
Disagrees with taxonomy
Split into >1 cluster
Differs from consensus
Unnamed species
Unnamed cluster
Named cluster
Singleton cluster
C
Genome 1
Genome 2
Genome n
Genome Cluster
Non-Redundant Pan Genome
Representative Genome
Phylogenetic Marker Genes


Per Sample
A Estimate Species Abundance
Metagenome
Phylogenetic Marker Genes
Map reads using HS-BLASTN
Taxonomic Abundance Profile
B Predict Gene Presence/Absence In Abundant Species
Pan Genomes
Build Pan Genome DB
Metagenome
Map reads using Bowtie2
C Identify Single-Nucleotide Polymorphisms in Abundant Species
Build Representative Genome DB
Representative Genomes
Map reads using Bowtie2
Metagenome
D
Copy Number Matrix
CNVs
Genomes
SNVs
Allele Frequency Matrix
Per species core genes
Per species CNV trees
Per species SNV trees
Per species core sites
Multiple Samples


A
Percent Unknown Species
0
20
40
60
80
100
Host Country
United States (114)
China (70)
Austria (62)
Denmark (75)
Spain (70)
Sweden (43)
Peru (72)
Tanzania (27)
B
Genome−Cluster Prevalence
(# samples)
0
100
200
300
400
Bacteroides vulgatus (16)
Blautia wexlerae (3)
Eubacterium rectale (3)
Bacteroides uniformis (7)
Faecalibacterium cf. (1)
F. prausnitzii A2−165 (1)
F. prausnitzii M21/2 (2)
F. prausnitzii L2−6 (1)
Bacteroides rodentium (1)
Oscillibacter sp. ER4 (1)
Ruminococcus bromii (1)
Parabacteroides distasonis (15)
butyrate−producing bacterium (4)
Alistipes putredinis (1)
Eubacterium eligens (1)
Bacteroides ovatus (11)
Ruminococcus torques (1)
Coprococcus comes (1)
Roseburia inulinivorans (1)
Eubacterium hallii (1)
Alistipes shahii (1)
Bacteroides thetaiotaomicron (4)
Parabacteroides merdae (3)
Bacteroides caccae (2)
Bifidobacterium longum (41)
Bacteroides xylanisolvens (9)
Odoribacter splanchnicus (1)
Blautia sp. KLE 1732 (2)
Alistipes onderdonkii (2)
Ruminococcus bicirculans (1)


A
Nucleotide Diversity (π)
0
0.005
0.015
* Inter−sample
● Intra−sample
Faecalibacterium prausnitzii M21/2
Eubacterium rectale
Ruminococcus bromii
Ruminococcus bicirculans
Faecalibacterium prausnitzii L2−6
Faecalibacterium prausnitzii A2−165
Eubacterium eligens
Faecalibacterium cf.
Roseburia inulinivorans
Bacteroides vulgatus
Akkermansia muciniphila
Oscillibacter sp. ER4
Barnesiella intestinihominis
Bacteroides stercoris
butyrate−producing bacterium SSC/2
Bacteroides thetaiotaomicron
Parabacteroides distasonis
Alistipes onderdonkii
Blautia wexlerae
Alistipes putredinis
Bacteroides ovatus
Parabacteroides merdae
Alistipes shahii
Bacteroides uniformis
Bacteroides caccae
B
Faecalibacterium prausnitzii
1
10
100
1000
10000
Count SNPs
Between Samples
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
10
20
30
40
Minor Allele Frequency (%)
0.000
0.010
0.02
π
Ruminococcus bromii
Between Samples
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
10
20
30
40
Minor Allele Frequency (%)
0.000
0.010
0.02
π
C
Faecalibacterium prausnitzii
P−value = 6.2e−06
π_{within}
0.000
0.005
0.010
0.015
Ruminococcus bromii
P−value = 7.4e−15
π_{within}
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Austria
China
Denmark
Peru
Spain
Sweden
Tanzania
United States


A
Bacteroides ovatus pangenome
(N=15,203 non-redundant genes)
present
absent
Samples
(N=147)
B
Percent of genes shared between samples
90
80
70
60
50
40
30
20
0.0–0.5
0.5–1.0
1.0–1.5
1.5–2
>2.0
Percent core-genome nucleotide divergence between samples
C
Core Genes
−log10(p−value
0
5
10
translation
GO:0006412
Nitrogen metabolism
KEGG:00910
Alanine, aspartate and glutamate metabolism
KEGG:00250
structural constituent of ribosome
GO:0003735
ribosome
GO:0005840
Oxidative phosphorylation
KEGG:00190
Glyoxylate and dicarboxylate metabolism
KEGG:00630
Pyrimidine metabolism
KEGG:00240
Auxiliary Genes
−log10(p−value
0
8
16
hypothetical protein
FIG00638284
Amino sugar and nucleotide sugar metabolism
KEGG:00520
Galactose metabolism
KEGG:00052
mobile element
FIG01306568
UDP−glucose 4−epimerase activity
GO:0003978
DNA replication
GO:0006260
DNA primase activity
GO:0003896
Starch and sucrose metabolism
KEGG:00500
Absent Genes
−log10(p−value
0
3
6
Sphingolipid metabolism
KEGG:00600
Glycosphingolipid biosynthesis − ganglio series
KEGG:00604
hypothetical protein
FIG00638284
Starch and sucrose metabolism
KEGG:00500
Putative exported protein precursor
FIG00967246
Aldo−keto reductase
FIG00639501
ribokinase activity
GO:0004747
Ribokinase
EC:2.7.1.15
Functions Enriched in Gene Sets
(KEGG Pathways, FIGFams, Gene Ontology, & ECs)
D
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Core (N=1877)
Auxillary (N=12445)
Absent (N=881)
20
40
60
80
100
% Gene GC Content


A
P–value = 3.9e–43
R–squared = 0.91
F_{ST} (Auxillary Genes)
F_{ST} (Core Genome SNPs)
0.0
0.1
0.2
0.3
0.4
0.5
B
Eubacterium rectale Phylogeny
Clade I
Clade II
Clade III
Clade IV
Host Country
United States (114)
China (70)
Austria (62)
Denmark (75)
Spain (70)
Sweden (43)
Peru (72)
Tanzania (27)
Eubacterium rectale ATCC 33656
Eubacterium rectale M104/1
Eubacterium rectale DSM 17629
0.01
C
Gene Sharing
0.6
0.8
Clade
I
II
III
IV
Eubacterium rectale
Gene Sharing