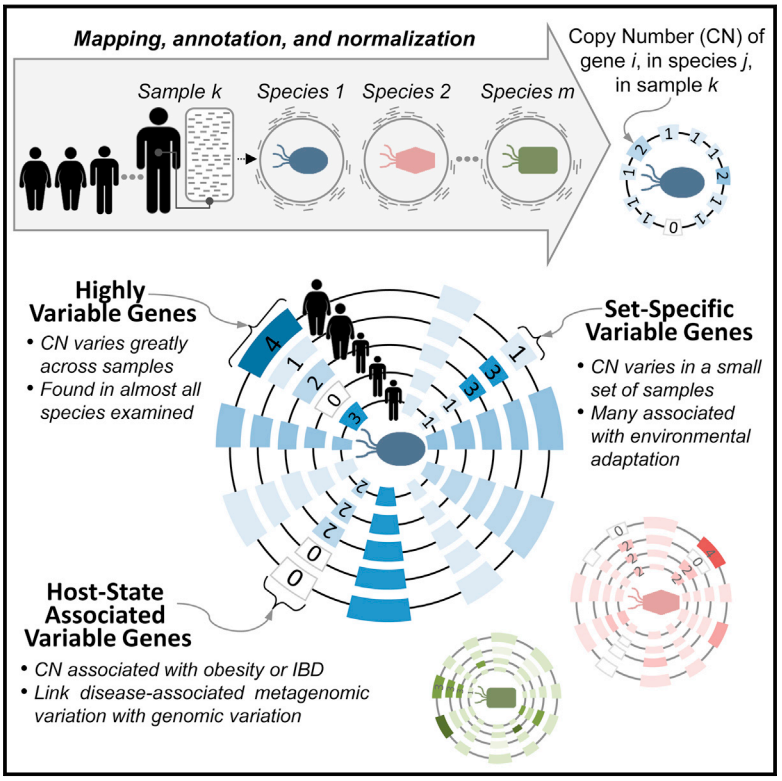
Article

Cell

Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species

Graphical Abstract



Authors

Sharon Greenblum, Rogan Carr, Elhanan Borenstein

Correspondence

elbo@uw.edu

In Brief

Extensive strain-level variation is detected in the human gut microbiome, with differences in gene copy-number impacting specific adaptive functions and linked to obesity and inflammatory bowel disease.

Highlights

- A metagenomic data analysis pipeline allows strain-level gene copy-number inference
- Copy-number variation (CNV) is widespread across many prevalent human gut species
- CNV involves mostly environment-related functions and is associated with disease
- Strain-level population structure reveals known and uncharacterized strains

Greenblum et al., 2015, Cell *160*, 583–594
February 12, 2015
http://dx.doi.org/10.1016/j.cell.2014.12.038

CellPress

Article

Cell

Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species

Sharon Greenblum,[1] Rogan Carr,[1] and Elhanan Borenstein[1,2,3,*]
[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[2]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
[3]Santa Fe Institute, Santa Fe, NM 87501, USA
*Correspondence: elbo@uw.edu
http://dx.doi.org/10.1016/j.cell.2014.12.038

SUMMARY

Within each bacterial species, different strains may vary in the set of genes they encode or in the copy number of these genes. Yet, taxonomic characterization of the human microbiota is often limited to the species level or to previously sequenced strains, and accordingly, the prevalence of intra-species variation, its functional role, and its relation to host health remain unclear. Here, we present a comprehensive large-scale analysis of intra-species copy-number variation in the gut microbiome, introducing a rigorous computational pipeline for detecting such variation directly from shotgun metagenomic data. We uncover a large set of variable genes in numerous species and demonstrate that this variation has significant functional and clinically relevant implications. We additionally infer intra-species compositional profiles, identifying population structure shifts and the presence of yet uncharacterized variants. Our results highlight the complex relationship between microbiome composition and functional capacity, linking metagenome-level compositional shifts to strain-level variation.

INTRODUCTION

The human gut microbiome plays an important role in host metabolism, immunity, and drug response and has a tremendous impact on our health (Iida et al., 2013; Kinross et al., 2011; Vijay-Kumar et al., 2010). Numerous comparative studies aiming to characterize the contribution of the microbiome to human health have already demonstrated marked shifts in the relative abundance of various species, genera, or phyla in various disease states (Frank et al., 2007; Hoffman et al., 2014; Larsen et al., 2010; Turnbaugh et al., 2009). Clearly, however, each microbial species represents many different strains that may encode considerably different sets of genes and a different number of copies of each gene (reflecting, for example, gene deletions and duplication events). Such intra-species variation endows each strain with potentially distinct functional capacities. Studies of individual isolates of cultured species have indicated, for example, that strains often differ in virulence (Gill et al., 2005; Salama et al., 2000; Solheim et al., 2009), motility (Zunino et al., 1994), nutrient utilization (Siezen et al., 2010), and drug resistance (Gill et al., 2005). Accordingly, the true functional potential of a microbiome cannot be inferred from species composition alone, and species-level comparative analyses may fail to capture important functional differences across samples. Recent efforts to catalog the relative abundance of known strains in human microbiome samples (Kraal et al., 2014) may recover some of these differences but are limited to sequenced reference genomes and are not able to identify novel, yet-to-be-sequenced variation. Gene-centric shotgun metagenomic studies, on the other hand, may identify genes or pathways that are differentially abundant across samples but cannot necessarily attribute these shifts to specific species or strains. Specifically, it is often unclear how much of the observed variation in gene composition is due to variation in the abundances of species and how much is contributed by intra-species variation. Indeed, conflicting results have been reported, with trends identified among species profiles that are often poorly translated to gene profiles and vice versa (Muegge et al., 2011; Turnbaugh et al., 2009). It is therefore not yet clear how prevalent gene-level intra-species variation is in the human gut, whether such variation is adaptive and affects specific functions, and how much of this variation has already been captured by reference genomes.

Some evidence already suggests that variation among strains is common in the human gut. Several studies have focused specifically on nucleotide-level variation, assessing, for example, the prevalence and stability of single-nucleotide polymorphisms across numerous metagenomes (Schloissnig et al., 2013) or the level of sequence diversity across multiple near-complete genomes from two bacterial species variants obtained by single-cell sequencing (Fitzsimons et al., 2013). Other studies have taken steps to associate sequence variation with gene-level differences, identifying, for example, areas of variable coverage and the coordinated loss of genes from specific gene families within the *Streptococcus mitis B6* genome (Human Microbiome Project Consortium, 2012) or a diverse array of strain-specific adhesion-like protein genes across cultured strains of *Methanobrevibacter smithii* (Hansen et al., 2011). Additional studies have used extensive manual genomic reconstruction to track strain-resolved shifts over time in *Actinomycetaceae* in the relatively low-complexity premature infant gut microbiome (Brown et al., 2013); to detect differences related to antibiotic resistance, transport, and biofilm formation among three strains of *Staphylococcus epidermis* (Sharon et al., 2013); or to identify

the variable presence of genes involved in transport, motility, carbohydrate metabolism, and virulence in two distinct strains of *Citrobacter* (Morowitz et al., 2011). These gene-level studies, however, mostly report small-scale or anecdotal results, focusing on one or a small number of species and often on specific gene families. A high-throughput, comprehensive analysis of gene-level variation across a large array of species in the human gut is therefore needed to more fully appreciate the extent and functional implications of strain variation in this complex microbiome.

To address this challenge, here we establish a rigorous and robust pipeline to estimate the copy number of each gene in a large set of prevalent gut microbial species in a given sample directly from metagenomic shotgun data and, furthermore, to detect copy-number variation across samples. We carefully calibrate this pipeline to confirm that it can successfully estimate the copy number of individual genes in individual species on a large scale. Applying this pipeline to 109 metagenomic samples from a recent study of the gut microbiomes of healthy, obese, and inflammatory bowel disease (IBD)-afflicted individuals, we estimate the copy number of more than 4,000 gene groups across 70 species in each of these samples and demonstrate the presence of widespread copy-number variation within many genes in many species. We find that specific functions are especially prone to copy-number variation, including functions relevant to a community lifestyle and adaptation to the gut environment, and further detect associations between strain variation and host phenotype. Finally, we demonstrate that these copy-number estimates can be used both to model the composition of known strains within each sample and to offer insight into complex population structures, suggesting the presence of yet uncharacterized species variants.

RESULTS

A Pipeline for Calculating Genomic Copy-Number Estimates in Metagenomic Samples

We developed a pipeline to confidently detect variation in gene content and gene copy number in a large set of prevalent human gut microbes directly from metagenomic data (Figure 1 and Experimental Procedures). Briefly, this pipeline works as follows. Shotgun metagenomic short reads were first aligned to a set of reference genomes representing dominant and prevalent gut microbiome strains. To account for the potentially multiple genomes available for each species in this reference database, genomes were grouped into clusters using a previously introduced sequence similarity-based method (Schloissnig et al., 2013). These clusters represent approximate species-level groups, though in some cases may not reflect classical taxonomic divisions. We used extensive simulations to carefully select alignment parameters and confirmed that, with these parameters, reads mapped to the correct region and correct genome cluster, whereas reads from genome clusters not represented in our reference database remained unmapped (Figure 2A; Figure S1; Extended Experimental Procedures). In parallel, gene coding regions from all reference genomes were annotated with KEGG orthology groups (KOs). Reference genomes and KOs with low confidence mapping were identified and excluded (Figure S2; Extended Experimental Procedures). For each sample, coverage across each KO-annotated region in each reference genome was calculated, and coverage values across regions corresponding to the same KO in the same genome cluster were summed. We then used the average coverage of 13 single copy marker genes, carefully selected for their universality, mapping accuracy, and coverage consistency (Figure S3; Extended Experimental Procedures), to convert the calculated coverage of each KO in each cluster to a copy-number estimate (Experimental Procedures). Overall, this process estimated the copy number, V_{kcs}, of each KO k, in each genome cluster c, detected in each sample s (Figure 1). Notably, copy-number estimates represent an average across the various genomes associated with each cluster that are present in the sample and across the potentially multiple genes associated with each KO. We further performed an analysis of an extensive synthetic dataset to confirm that this scheme accurately recovers species abundances and copy-number values (Figures S4A and S4B; Extended Experimental Procedures).

We applied this pipeline to a dataset of 109 previously collected gut metagenomic samples from a Danish/Spanish cohort (Qin et al., 2010), mapping in total >2.45 billion 75 bp reads to 235 reference genomes grouped into 96 genome clusters (Table S1; Extended Experimental Procedures). The average coverage across the 13 marker genes (a proxy for the abundance of each cluster in each sample) varied considerably across clusters and between samples (Figures 2B and 2C). To limit any downstream analysis to high-confidence copy-number estimates, we therefore considered only genome clusters with sufficient coverage in a sample (which we term "detectable" clusters; Experimental Procedures). We identified a total of 70 clusters that were detectable in at least one sample, with an average of 16 detectable clusters in each sample (Table S2). Overall, this analysis assigned copy number values to ~1.5 million KO-cluster-sample triplets, estimating the copy number of thousands of KOs across a large array of genome clusters in >100 samples (Table S3).

This dataset of copy-number estimates provides a first large-scale account of gene-level strain variation among organisms common to the human gut. Below, we mine this dataset to explore neutral and adaptive variation in this highly complex ecosystem in a manner that goes beyond species-level comparative analysis. Importantly, this dataset and the pipeline described above can serve as a valuable resource for future studies of compositional shifts in the human microbiome and in other environments, linking metagenome-level differences in gene abundance to genome-level variation.

Identifying Genes with Highly Variable and with Set-Specific Variable Copy Number

Given the copy-number estimates obtained above, we set out to identify specific KOs in specific clusters (KO-cluster pairs, or KCs) whose copy number varied across samples. Notably, to detect variation, we compared the copy number of each KC across different samples rather than comparing the estimated copy number in any given sample to the copy number in a reference genome, avoiding spurious variation predictions that may result from annotation errors or bias in the set of reference genomes. Clearly, many clusters can be detected in only a few samples. To confidently detect copy-number variation, we

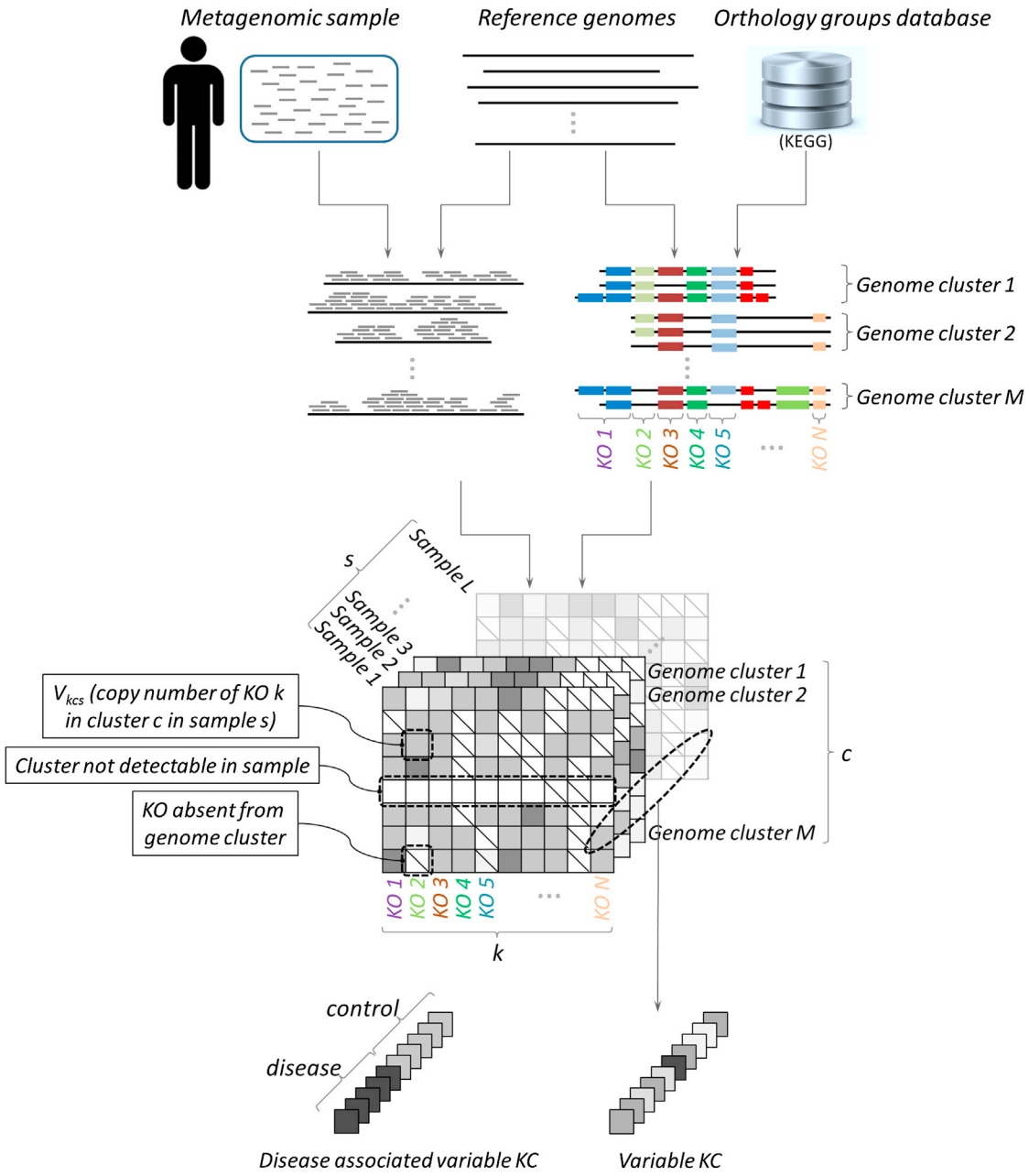


Figure 1. Schematic of Analysis Pipeline

Reads from metagenomic samples were mapped to KEGG-annotated reference genomes, grouped into species-level genome clusters. The total coverage of each KO (KEGG orthology group), *k*, in each genome cluster, *c*, in each sample, *s*, was normalized by cluster abundance to calculate gene copy number V_{kcs}. KCs (specific KOs in specific genome clusters) whose copy number varied significantly across samples were detected, as well as those whose copy number was associated with host state (obesity, IBD).

See also Figure S3 and Table S3.

therefore only considered the 40 clusters that were detectable in at least 10 samples.

We first set out to identify KCs that exhibit extreme and prevalent variation across samples. Specifically, we calculated the level of inter-sample variation in the copy number of each KC and defined as *highly variable* those KCs whose variation was at least two standard deviations greater than the average variation of all KCs (Experimental Procedures). We used both cross-validation analysis and synthetic samples to confirm the robustness and accuracy of this approach (Extended Experimental Procedures; Figure S4C). In total, this analysis detected 735 highly variable KCs spanning 261 KOs across 38 genome clusters (Figure 3; Table S4). The number of highly variable KCs in each cluster varied greatly, reaching up to 47 KCs in

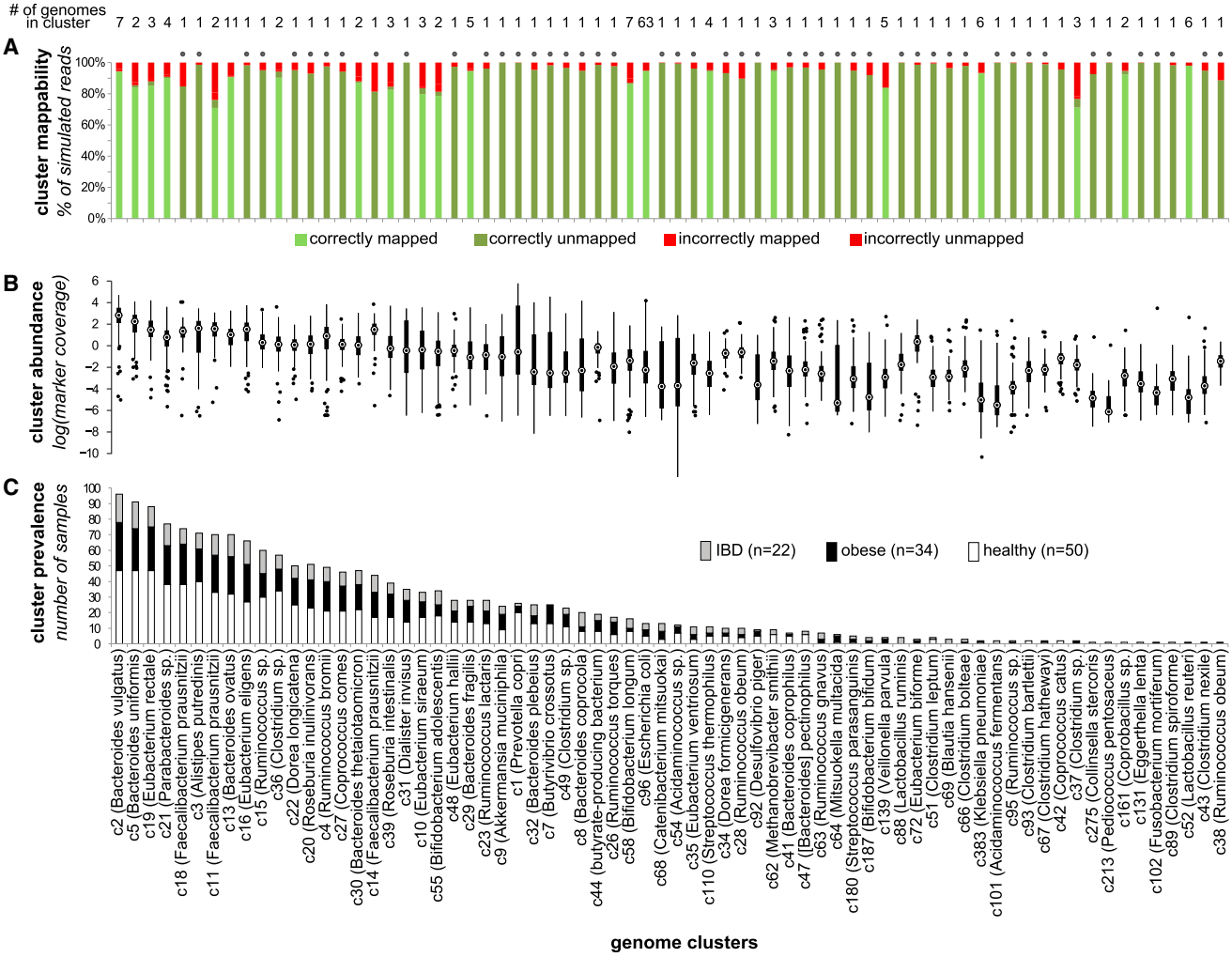


Figure 2. Genome Cluster Statistics

The mappability, abundance, and prevalence of each genome cluster (representing a species-level group) are shown in three vertically aligned plots. Clusters are sorted by their prevalence across samples.

(A) Cluster mappability, as determined by a large-scale simulation assay measuring the accuracy of mapping reads extracted from the cluster's genomes to a database in which the genome of origin was removed. In this simulation, reads from clusters represented in the reference database by a single genome (marked with a dot above the column) are expected to remain unmapped.

(B) The distribution of each cluster's abundance across samples, as determined by the average coverage of 13 single-copy marker genes.

(C) Cluster prevalence (the number of samples in which the cluster was "detectable") within each host group, shown as a stacked bar plot.

See also Figures S1 and S2 and Tables S1 and S2.

the *Roseburia intestinalis* cluster (representing 4.05% of the KCs in this cluster), with an average of 1.79% of the KCs in each cluster (Table S5). We found no apparent relationship between the amount of variation observed in a cluster and the number of reference genomes in the cluster or the prevalence of the cluster across samples, but we did observe a tendency toward high variation in species from the *Firmicutes* phylum compared to other species (t test, $p < 0.05$; see also Figure 3). Although the majority of highly variable KOs (57.1%) were variable in just one cluster, certain KOs were variable across many clusters, with some KOs variable in ten or more different clusters.

The analysis above focused on KCs that exhibit extreme variation and on KCs that vary greatly across *many* different samples. Variation within other genes, however, may be more subtle and may reflect, for example, adaptive variation that can be observed in only a small set of samples. We therefore set out to additionally identify *set-specific variable KCs*, wherein the copy number of a given KC was relatively constant across most samples but deviated significantly in a small subset of the samples (Experimental Procedures). In this analysis, we further distinguished cases in which a KC exhibited a consistently high copy number in this subset of samples compared to all other samples (set-specific *increased* copy number) from cases in which a KC exhibited a consistently low copy number in this subset of samples (set-specific *decreased* copy number) or in which it exhibited increased copy number in one subset and decreased in another. As

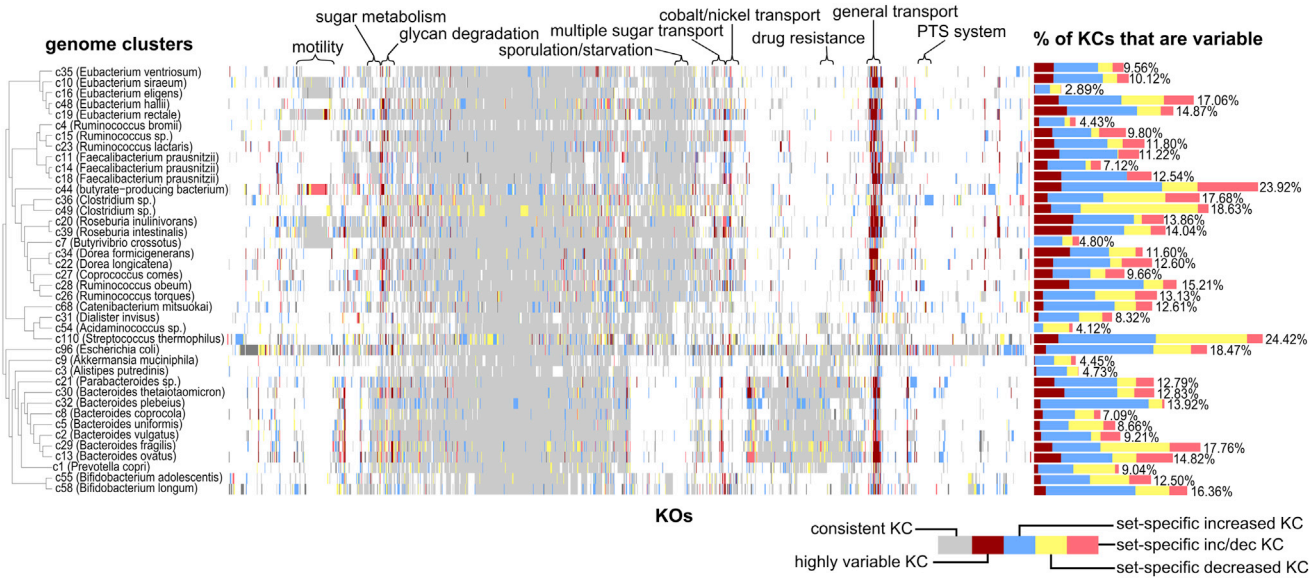


Figure 3. A Map of Variable KCs

A matrix map representing the status of variable KOs (x axis) in each genome cluster (y axis). Colored bars represent variable KCs (highly variable KCs vary widely in copy number across all samples, whereas set-specific variable KCs are increased and/or decreased in copy number in only a small subset of the samples), while light gray bars indicate KCs with consistent copy number across samples, and KOs not present in a genome cluster are left white. Genome clusters are ordered by phylogeny, and KOs are ordered by hierarchical clustering. The bar chart to the right of the map represents the fraction of KOs in each cluster identified as variable. Above the map, certain groups of functionally related KOs are highlighted. The 314 KOs uniquely variable in the *E. coli* cluster (the majority of which have only been annotated in *E. coli*-like genomes) were excluded due to space constraints.

See also Figure S4 and Tables S4–S6.

expected, we found that set-specific variable KCs were much more common than highly variable KCs. In total, our analysis detected 5,004 set-specific variable KCs covering 1,859 KOs across the 40 genome clusters examined (Figure 3; Table S4). In general, we observed more cases of set-specific increased copy number than of set-specific decreased copy number, but this ratio shifted markedly across clusters, and in certain clusters (i.e., *Clostridium sp.*, *Streptococcus thermophilus*) mostly set-specific decreased KCs were observed.

Detected Variation Captures Both Known and Novel Strain Variation

As validation of our pipeline and results, we compared the set of highly variable KCs obtained for each cluster to known variation among the cluster's sequenced reference genomes. Clearly, the reference genomes in our database do not capture the full extent of intra-species variation in the gut microbiome. Similarly, our samples likely do not include much of the variation present in our reference genomes, as many of these reference genomes represent strains isolated from clinically distinct individuals, phenotypically diverse cohorts, or non-gut samples. Accordingly, a large number of genes that vary in copy number across reference genomes may still exhibit consistent copy number across the gut samples analyzed above. Yet, the set of detected highly variable genes, which aims to include genes that vary frequently in their copy number across genomes, is likely to capture many instances of known variation in gene content among available reference genomes. Indeed, considering the 15 multiple-genome clusters in our database, a striking 81% of the detected highly variable KCs also vary in copy number across reference genomes (Figure 4). Moreover, in seven of these clusters, *all* highly variable KCs also vary in copy number across reference genomes. Notably, six of these clusters contain at least three genomes, whereas the majority of the other clusters contain only two, suggesting that more sequenced strains may be needed to fully capture the variation associated with these clusters (and more importantly, with clusters for which only a single genome was available). Importantly, we demonstrated that a similar overlap can be observed when comparing predicted variation to known variation among a large collection of genomes *not* included in our database, confirming that this overlap is not an artifact of the specific reference genomes used in our analysis (Figures 4B and 4C; Extended Experimental Procedures). Comparison of set-specific variable KCs to known variation across reference genomes again confirmed that the variation detected greatly overlapped with known variation observed across sequenced strains (Figure S5). Notably, however, set-specific variable KCs also included many instances of novel variation, suggesting that the set of reference genomes currently available does not capture the full extent of copy-number variation in the gut. Comparison of detected set-specific variation to variation observed across two manually assembled *Citrobacter* strains further revealed significant overlap (Extended Experimental Procedures).

Functions Associated with Variable Genes

We examined whether the detected copy-number variation was associated with specific functions in each genome cluster. We first used enrichment analysis to identify functions that were

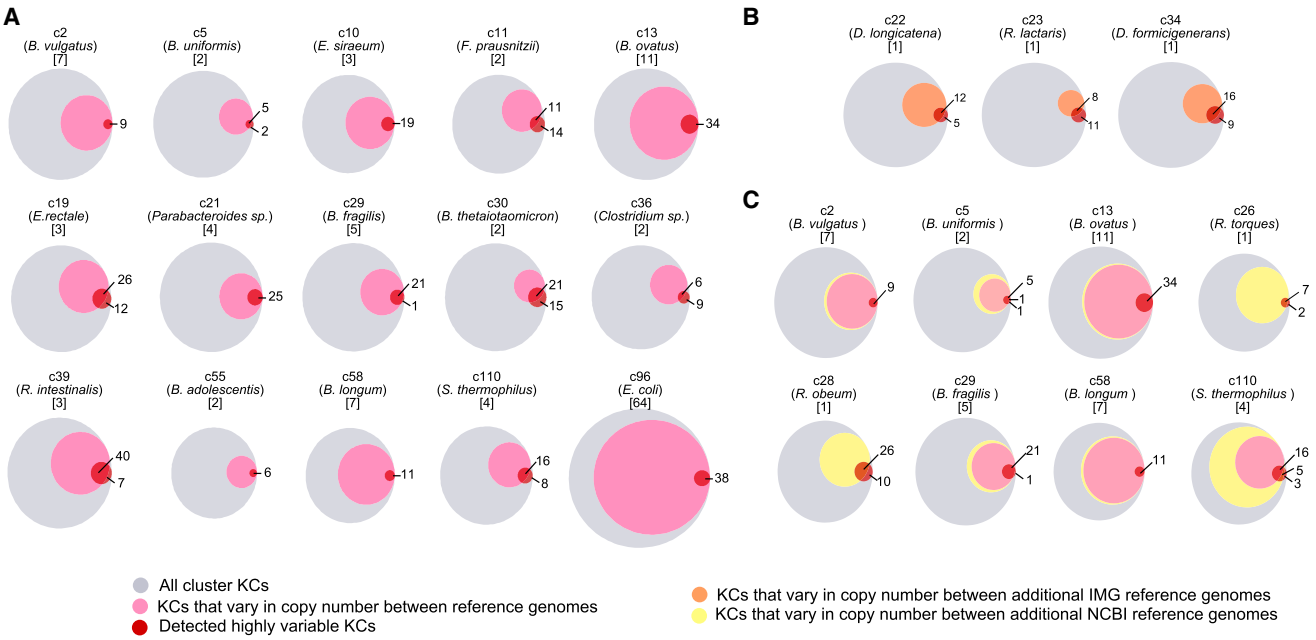


Figure 4. Comparison of Highly Variable KCs to Known Variation among Reference Genomes

(A) In each Venn diagram, the gray circle represents the set of all KCs in a given genome cluster, the pink circle represents the fraction of those KCs exhibiting copy-number variation across the cluster's reference genomes, and the red circle represents the set of KCs detected as highly variable. Overlap of the pink and red circles indicates correspondence between known and detected variation. Each diagram is labeled with the cluster ID, representative species name, and number of reference genomes.

(B and C) Additional variation in reference genomes that were not used as mapping targets is represented by either an orange circle (additional reference genomes from IMG) or a yellow circle (additional reference genomes from NCBI), compared to variation in included reference genomes (pink) and detected highly variable KCs (red).

See also Figure S5.

over-represented among the set of highly variable KCs in each cluster. We found that transport-related functions were overwhelmingly prone to high copy-number variation (Table S6). Specifically, ten of the genome clusters analyzed were enriched for variation in KCs associated with transport annotations, including the general BRITE term "Transporter," as well as more specific modules related to either sugar or iron complex transport. For example, within the *Bacteroides ovatus* cluster, seven of the cluster's 66 transport-associated KCs were highly variable (Figure 5), including all three KCs (K02013, K02015, K02016) involved in a specific iron complex transport system module (M00240). Interestingly, significant variation in sugar transport functions was only found among clusters in the phyla *Firmicutes* and *Actinobacteria*, whereas *Bacteroidetes* clusters were uniquely associated with variation in the iron complex transport system (see Table S6). Studies of cultured organisms from various environments and experimental evolution assays have suggested that loss, amplification, and acquisition of transport functions constitute a primary adaptive mechanism (Gevers et al., 2004; Heikkinen et al., 2007; Lee and Marx, 2012; Sonti and Roth, 1989); here, we show that this flexibility in the copy number of transport genes likely extends to a considerable proportion of prevalent gut species and that, within this general class, specific transport genes may facilitate adaptation to the gut environment.

We additionally found that motility-related KCs were highly variable in the *Eubacterium rectale* genome cluster. Specifically, in this cluster, 7 of the 38 highly variable KCs were bacterial motility proteins, of which four were structural flagellar components, two were involved in chemotaxis, and one was essential for twitching motility (Han et al., 2008). Motility proteins, and especially flagellar proteins, are widely associated with virulence and immunostimulation, and the gain or loss of flagellar components is believed to be an important adaptive mechanism (Borziak et al., 2013; Heikkinen et al., 2007; Al Mamun et al., 1997). Moreover, variation in these seven KCs was highly consistent within samples; most samples contained either detectable copies of all seven KCs or no (or low number of) copies of all of these KCs (Figure S6). Though we found no variation in the copy number of any of these genes among the three sequenced reference genomes included in the *Eubacterium rectale* cluster in our database, a recent study of 27 elderly gut metagenomes identified non-uniform coverage of genes involved in the flagellum biogenesis pathways of six *Eubacterium* and *Roseburia* species (Neville et al., 2013), suggesting that the current reference genomes may not capture the full dynamic range of these species.

Next, we considered the collection of set-specific variable KCs and examined their functional annotations. Interestingly, hierarchical clustering of set-specific variable KOs based solely on their variation profile across the 40 clusters revealed distinct groups of functionally related genes that vary in a given genome cluster or within multiple clusters (Figure 3). For example, a large set of genes related to cell growth and

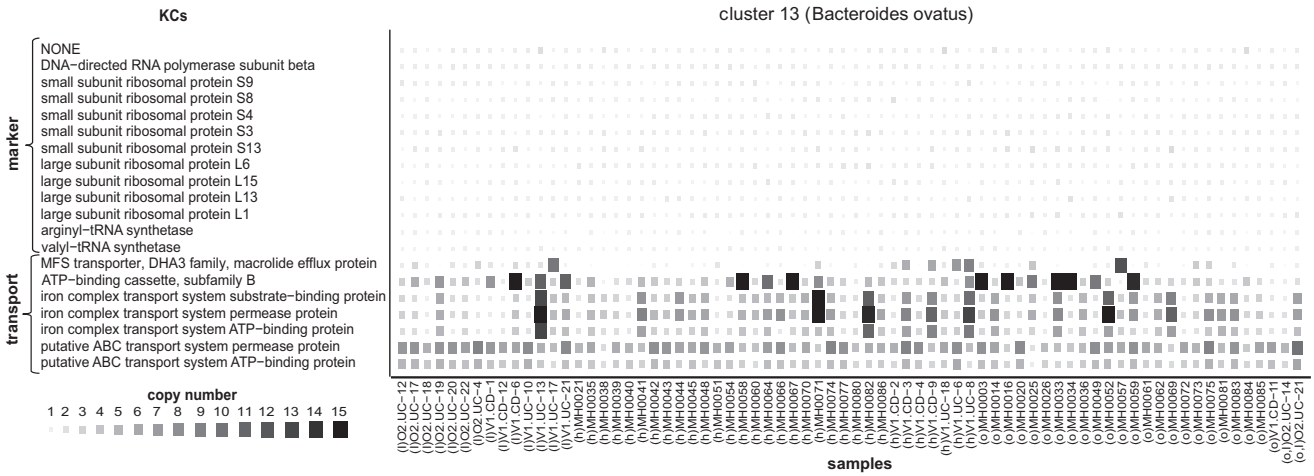


Figure 5. Copy Number of Highly Variable Transport KCs in *Bacteroides ovatus*

The size and color of each square represent the copy number of each highly variable KC within each sample. Samples are grouped by host state (I, IBD; h, healthy; o, obese). The copy numbers of the 13 marker KCs in this genome cluster are illustrated for comparison. See also Figure S6.

sporulation were all identified as set-specific variable KCs in the two genome clusters associated with *Clostridium sp*. Similarly, a set of sugar metabolism genes were all identified as set-specific variable KCs in *Roseburia intestinalis*, and a number of antibiotic resistance genes were identified as variable in multiple genome clusters, primarily those in the *Firmicutes* phylum. An enrichment analysis of functions associated with set-specific variable KCs in each cluster additionally revealed a number of important functions that were prone to copy-number variation (Table S6). For example, genes in the lipopolysaccharide biosynthesis pathway in *Dialister invisus* and *Clostridium sp.* were often observed with a higher copy number in a small set of samples. Interestingly, variation within functions related to sugar metabolism (i.e., KEGG pathways galactose metabolism, starch and sucrose metabolism, fructose and mannose metabolism, polyketide sugar unit biosynthesis) was observed primarily within *Bacteroidetes* clusters, whereas set-specific transport-related variation was almost absent from these clusters. Other functions enriched for set-specific variable KCs suggest transitions between virulent states, such as motility in *butyrate-producing bacteria* (NCBI accession FP929062), *Eubacterium rectale*, and *Clostridium sp.*; streptomycin biosynthesis in *Acidaminococcus sp*.; lysosyme production in *Bacteroides ovatus*; the EHEC/EPEC pathogenicity signature in *Escherichia coli*; and secretion systems in *butyrate-producing bacteria* (NCBI accession FP929062), *Clostridium sp*., and *Escherichia coli*. Within *Escherichia coli*, type II secretion system genes were identified as set-specific decreased copy-number KCs, whereas type III secretion system genes were identified as set-specific increased copy-number KCs. Overall, much of the observed variation appeared to be associated with the way a species responds to and interacts with its surroundings, highlighting the strong adaptive potential of gut-associated bacteria.

Clearly, different cohorts could harbor different sets of strains owing to an assortment of ecological or host-specific factors, and accordingly different genes may vary in copy number in different datasets. Notably, however, analysis of a second dataset of 73 gut samples from a Chinese cohort (Qin et al., 2012) yielded a marked overlap with our original Danish/Spanish cohort in both the set of KCs identified as variable and the set of functions enriched for copy-number variation (Extended Experimental Procedures). These findings suggest that, although variation may be personal, certain genes and functions (e.g., those related to environmental adaptation) may be universally prone to variation.

Host State-Associated Variation

Although much of the variation across strains may reflect neutral processes or transitory dynamics, some variation may represent adaptation to a specific host phenotype. To detect such potentially adaptive variation, we identified variable KCs in which the copy number in samples from obese or IBD subjects was significantly different than in samples from healthy subjects (Experimental Procedures). In total, we found 24 KCs whose copy number was significantly associated with IBD and three KCs whose copy number was significantly associated with obesity (FDR < 0.05; Table S7).

Interestingly, a number of these KCs have been previously implicated in adverse host health states. For example, in our analysis, obesity was associated with a higher copy number of thioredoxin 1 (K03671) in *Clostridium sp.* (Figure 6A), and indeed thioredoxin reductase was recently shown to be enriched in the cecal metaproteome of mice fed a high-fat diet (Daniel et al., 2014). Such results are consistent with thioredoxin's regulatory role in maintaining redox equilibrium and the demonstrated links between a high-fat diet and oxidative stress in mammals (Furukawa et al., 2004). Additionally, in our analysis, the loss of a ubiquinone-reducing gene (K00349; *nqrD*) from *Bacteroides plebeius* was associated with obesity. A recent study in mice showed that supplemental ubiquinone reduced inflammation and metabolic stress accompanying a high-fat high-fructose diet by reducing the expression of certain genes associated

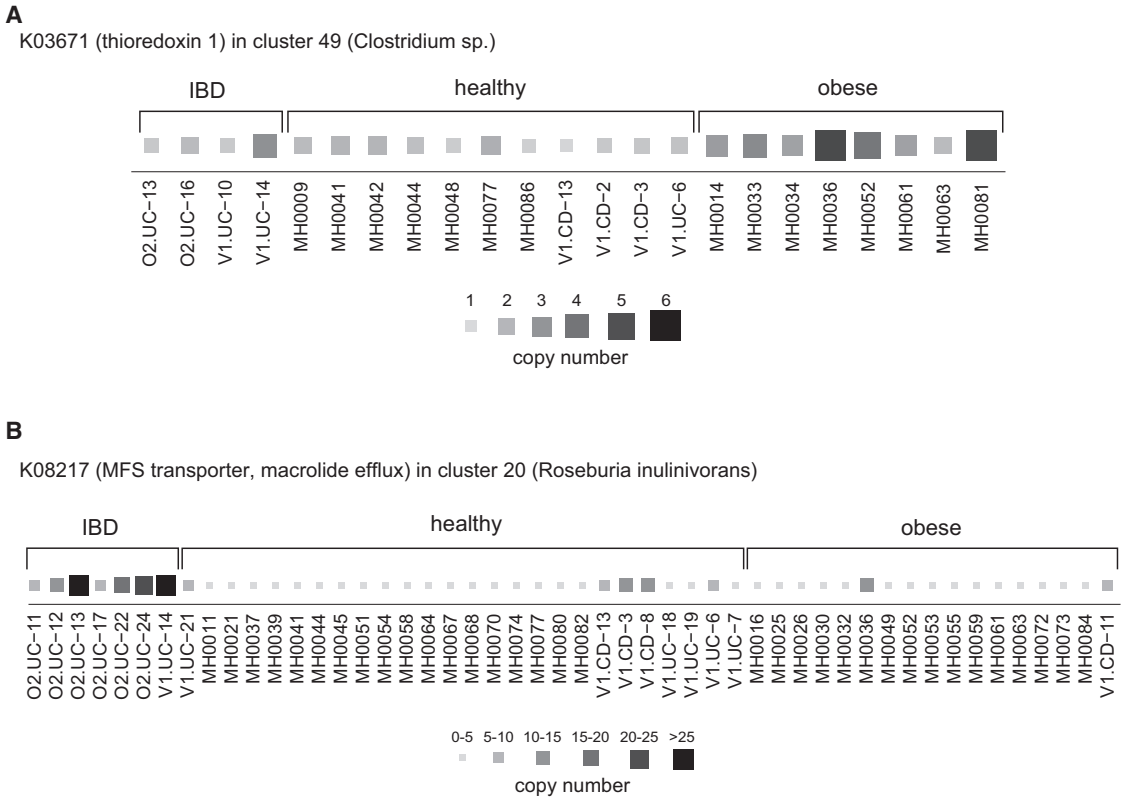


Figure 6. Copy-Number Variation of Host State-Associated KCs

Two KCs whose copy number was significantly increased in samples from a specific host state are shown. The size and color of each square represent the copy number of the KC within each sample.

(A) The copy number of thioredoxin 1 (K03671) in Clostridium sp. is significantly increased in samples from obese subjects.

(B) The copy number of an MFS transporter gene (K08217) in the Roseburia inulinivorans genome cluster is significantly increased in samples from IBD subjects. See also Table S7.

with stress-response (Sohet et al., 2009), while mice not receiving the supplement gained more weight than their counterparts. Importantly, however, ubiquinol, the reduced form of ubiquinone, has recently been shown to be the more readily absorbed and more active form of the compound (Langsjoen and Langsjoen, 2014), raising the possibility that loss of microbial ubiquinone-reducing capabilities from certain species may hinder the effectiveness and protective capacity of ubiquinone in the host. Other findings shed new light on the role of individual species in disease, with evidence of variation associated with common disease hallmarks, such as pathogenicity-related secretion and antibiotic resistance. In *Roseburia inulinivorans* (Figure 6B), increased copy number of a gene (K08217) coding for a major drug efflux protein known to play a role in antibiotic resistance was highly associated with IBD-afflicted individuals. Similarly, HlyD (K01993), an essential component of RTX hemolytic toxin secretion (Pimenta et al., 2005), exhibited increased copy number in IBD samples in *Bacteroides uniformis.* See Table S7 for a full list of disease-associated KCs. Interestingly, none of the obesity-associated KCs and only 3 of the 24 IBD-associated KCs were found to vary significantly in the Chinese cohort described above, among whom only one individual was obese and none were reported as having IBD.

Strain-Level Deconvolution of Microbiome Composition and Intra-Species Population Structure

Clearly, the microbiomes of different individuals can house multiple strains of the same species with potentially different relative abundances. Our copy-number estimates for each cluster accordingly represent average copy numbers across the different strains in the sample. Next, we therefore examined whether these estimates can be used to obtain insights into strain-level population structure, going beyond species-level composition assays and focusing specifically on the composition of strains within each genome cluster rather than on the abundance of the cluster itself.

First, we explored how well the copy-number profiles obtained for each genome cluster in each sample can be explained by known reference strains, using a regression analysis to deconvolve these copy-number profiles into a linear combination of the strains included in our database (Experimental Procedures). Obviously, these strains may not encompass the full set of strains present in the samples analyzed, yet such an analysis may be useful in examining what portion of the observed variation can be accounted for by known strains and what portion represents potentially novel variation. Indeed, we found that, in well-characterized clusters with many sequenced genomes,

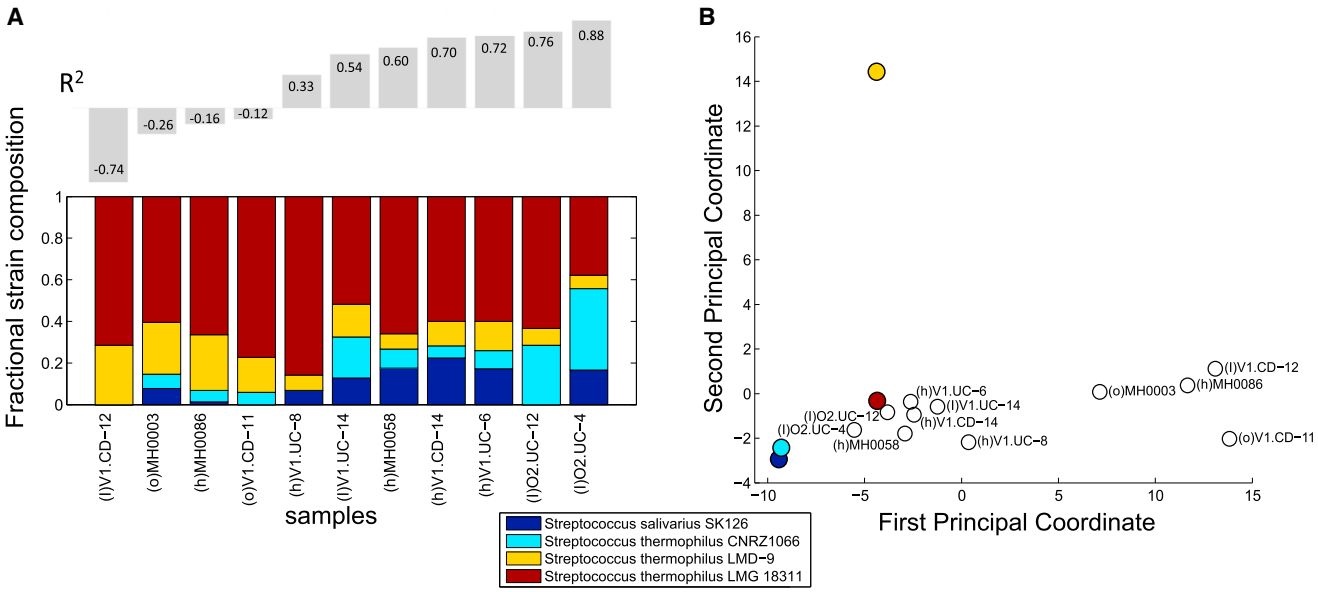


Figure 7. Predicted Strain-Level Population Structure within *Clostridium sp*

(A) A linear regression analysis was used to model the copy-number profile obtained for cluster 110 (*Streptococcus thermophilus*) in each sample as a combination of known reference genomes, with prediction weights shown as stacked colored bars. Prediction accuracy (R^2) is indicated above each bar. Samples with low or negative R^2 values potentially contain variation that cannot be explained by any combination of known reference genomes.

(B) A principal coordinate analysis depicting the differences between the copy-number profiles obtained for this genome cluster in the various samples (open circles), as well as the copy-number profiles of reference genomes (filled circles).

See also Figure S7.

the copy-number profiles of most samples could be well explained by a linear combination of known strains. For example, in the *Escherichia coli* cluster that comprised 63 sequenced genomes in our database, 76% of the variation in copy number could be explained on average by these genomes ($R^2 = 0.76 \pm 0.12$). In this cluster, the inferred representation of each strain differed widely across samples, with some strains (i.e., *Escherichia coli O111:H- str. 11128*) highly represented across multiple samples and others found in just one sample. However, for less well-characterized clusters with only a few known strains in our database, in some cases just a subset of the observed copy-number variation could be explained. For example, the four known strains of *Streptococcus thermophilus* could be used to explain a majority of the variation observed in some of the samples ($R^2 > 0.5$) yet failed to explain the variation observed in four of the samples ($R^2 < 0$), suggesting the existence of potentially novel, yet-to-be-sequenced variation (Figure 7A).

To further compare copy-number variation profiles across samples and to examine variation that may not be captured by known strains (including notably, in clusters comprising only one known strain), we used a principal coordinate analysis. This analysis revealed a complex population structure within each cluster, with marked differences among samples indicating the prevalence of personalized variation. For a number of genome clusters, however, samples appear to group into distinct sets, potentially reflecting individuals with similar intra-species population structures (Figure 7B). Moreover, by including the reference genomes in this principal coordinate analysis, we were able to distinguish previously captured variation versus novel variation observed across samples. For example, the principal coordinate plot for the *Streptococcus thermophilus* genome cluster (Figure 7B) clearly demonstrates that, although the copy-number profiles of most samples clustered tightly with several known reference genomes, the four poorly explained samples mentioned above clustered together and contained variation that was distinct from any reference genome. Such a pattern may indicate the presence of novel shared strains, providing a promising basis for targeted sequencing. Similar patterns were also observed in other clusters, in which a distinct, tightly clustered subset of samples or individual samples exhibit markedly different copy-number profile from that of any sequenced genome (Figures S7A and S7B). Overall though, each genome cluster exhibited a unique population structure across individuals, highlighting the complex suite of forces governing taxonomic composition in the gut (Levy and Borenstein, 2013).

DISCUSSION

By and large, closely related organisms tend to encode similar sets of genes. This consistency is in fact often used to infer functional capacity from taxonomy (Langille et al., 2013; Zaneveld et al., 2010). Clearly, however, this relationship between phylogeny and gene content is imperfect, and each species represents a large collection of strains that differ in the set of genes they encode, the copy number of these genes, and ultimately, their functional capacity. Above, we have focused on identifying instances in which this relationship between microbial species

and genes breaks, presenting a large-scale analysis of copy-number variation in a diverse array of gut species. Our analysis has demonstrated that copy-number variation is prevalent in the gut environment, with some species exhibiting significant copy-number variation in >20% of their genes. Such variation may induce significant microbiome-wide shifts and may account for at least some of the observed discrepancies between trends observed at the species levels versus trends measured at the gene level. Moreover, intra-species variation was shown to be especially prevalent in genes involved in specific functions, most notably functions that impact the way an organism interacts with its environment such as transport and signaling processes. This may suggest an adaptive dynamic by which certain species respond to changes in community composition or in the gut niche and a potentially crucial role of the gut environment in shaping bacterial evolution (Levy and Borenstein, 2013; Shapiro et al., 2012). Other highly variable functions, such as lipopolysaccharide biosynthesis, cell motility, and secretion systems, may represent changes in virulence as organisms respond to host immune responses. Interestingly, many of these same functions were highlighted in a previous study as more difficult to accurately correlate with 16s data (Langille et al., 2013). Our analysis further identified variable functions that may correlate with host states, exhibiting differential copy number in specific genomes. It remains unclear, however, whether such host state-associated variation is a cause or an effect. Our framework additionally facilitated the inference of intra-species population profiles for each individual, suggesting that most individuals harbor multiple strains of each species.

Though still far from an exhaustive catalog of strains that may be present across all human gut microbiomes, the framework presented above represents the most comprehensive account of copy-number variation in the human gut microbiome to date. It is our hope that this framework and the results presented here will inform future studies of strain-level microbiome composition, demonstrating the extent of functional information that is lost by limiting characterization to the level of species and prompting further investigation and sequencing of strain-level features. Yet, there are clearly a number of caveats that should be considered in designing such future efforts. First, our analysis is limited to the detection of variation in gut species for which at least one fully sequenced genome is available, and future studies may benefit from additional genomes. Notably though, we did not detect significantly more variation in clusters for which more reference genomes were available. In addition, our pipeline was designed to detect gene losses or amplifications but cannot identify gain of genes that are not present in any of the reference genomes included in the genome cluster. Such gain or transfer events may represent an additional substantial source of intra-species variation (Smillie et al., 2011). Our framework could, however, further facilitate future efforts to study sequence divergence among duplicated genes, informing our view of neofunctionalization and conservation processes in the microbiome. Notably, in our analysis, we focused on detecting high-confidence instances of variation, applying conservative parameters for read alignment and for variability calling. Specifically, we limit our analysis to "detectable" genome clusters, defined as those with >1× coverage in the sample. Our analysis of a synthetic dataset confirmed that, in such clusters, copy-number estimates can be inferred with 96% accuracy but that prediction accuracy dropped significantly in genome clusters with lower coverage (Figure S4B and Extended Experimental Procedures). With 13 million reads per sample (the lowest sequencing depth in the cohort analyzed), species that comprise >0.4% of the sample are likely to be considered detectable by our pipeline (while a higher sequencing depth of a sample will clearly allow analysis of even rarer species). Future studies may relax some of these parameters or incorporate additional information (e.g., gene conservation) to detect more subtle variation. Finally, as with most studies relating microbiome composition to function, our analysis relies on the availability of functional gene databases, which may contain incomplete or erroneous annotations. By considering variation across samples rather than variation from reference genomes, our analysis is largely robust to such annotation inaccuracies. Interestingly, however, variable KCs identified by our analysis were much more likely to lack a functional annotation than non-variable KCs, suggesting that much of the detected variation in gene content has as yet uncharacterized consequences. Combined, these results highlight both the need for additional genome sequences and the importance of continued efforts for characterizing gene function.

Ultimately, analysis of intra-species variation in microbial communities is crucial for understanding the complex relationship between species composition and community-level functional capacity. Our analysis, quantifiably characterizing such variation in the gut microbiome, is an important first step in this direction, and the resulting dataset provides an essential resource for future predictive studies.

EXPERIMENTAL PROCEDURES

Metagenomic Samples and Reference Genomes

Gut metagenomic data for 109 Danish and Spanish individuals, including individuals afflicted with obesity or IBD, was obtained from (Qin et al., 2010). A list of 261 dominant and prevalent human gut microbial strains, grouped into 101 genome clusters (Table S1) based on sequence similarity of 40 marker genes, was obtained from (Schloissnig et al., 2013). Nucleotide contig sequences, gene calls, and amino acid protein sequences were downloaded for each genome, and protein sequences were annotated with KEGG orthologous groups (KOs). See Extended Experimental Procedures for more details.

Calculation of Copy-Number Estimates

Shotgun metagenomic reads were aligned to the set of reference genomes with BWA, using parameters and filters carefully validated by extensive simulation analyses (Figures S1 and S2; Extended Experimental Procedures). In total, 2,469,102,286 reads were mapped. Average coverage over each gene region was determined using samtools (Li et al., 2009), and the coverage of each KC (KO-cluster pair) was obtained by summing over all genes annotated with the same KO and genome cluster. KC coverage was normalized by cluster abundance, defined as the average coverage over a set of 13 universal marker KOs (Figure S3B; Extended Experimental Procedures), to obtain the estimated copy number V_{kcs} of each KO k, in each cluster c, and in each sample s. "Detectable KCs" in a sample were defined as those with $V_{kcs} >= 0.5$. "Detectable clusters" within each sample were defined as those with at least 12 detectable marker KCs and average marker coverage ≥ 1. KCs that were not detectable in any sample were removed from the analysis.

Detection of Highly Variable and Set-Specific Variable KCs

For each of the 40,088 KCs present in clusters detectable in at least ten samples, the median copy number (baseline) across samples and the MAD

(median absolute deviation) from this baseline were calculated. KCs with a MAD more than 2 SDs from the MAD distribution mean (MAD > 0.6346) were considered *highly variable*. KCs in which at least 10% of samples had a copy number that exceeded the baseline by this threshold were considered *set-specific increased variable KCs*. *Set-specific decreased KCs* were similarly defined as KCs in which at least 10% of samples had a copy number that fell below the baseline by this threshold.

Detection of Host State-Associated KCs

A KC was defined as obesity associated if the copy numbers in samples from obese individuals were significantly higher or significantly lower than the copy numbers in samples from non-obese individuals, according to a two-sample t test (FDR-corrected $p < 0.05$). IBD-associated KCs were similarly defined. Samples that were labeled as both obese and IBD were omitted from this analysis.

Copy-Number Profile Deconvolution and Principal Coordinate Analysis

For each sample, a non-negative least-squares linear regression analysis was performed to obtain the linear combination of reference genomes in each multi-genome cluster, optimally explaining the copy-number estimates of variable KCs. The regression was constrained such that the sum of genome weights for each sample and cluster equaled one. Prediction error was defined as the R^2 value for each sample. A principal coordinate analysis was also performed for every genome cluster, operating on the pairwise Euclidian distance matrix of set-specific variable KC copy numbers in each sample and each sequenced reference genome.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.12.038.

ACKNOWLEDGMENTS

We thank Peter Turnbaugh, Jay Shendure, Phil Green, Colin Manoil, two anonymous reviewers, and the members of the Borenstein Lab for support and helpful discussions. This work was supported by a New Innovator Award DP2 AT 007802-01 to EB.

Received: June 19, 2014
Revised: October 30, 2014
Accepted: December 24, 2014
Published: January 29, 2015

REFERENCES

Al Mamun, A.A., Tominaga, A., and Enomoto, M. (1997). Cloning and characterization of the region III flagellar operons of the four Shigella subgroups: genetic defects that cause loss of flagella of Shigella boydii and Shigella sonnei. J. Bacteriol. *179*, 4493–4500.

Borziak, K., Fleetwood, A.D., and Zhulin, I.B. (2013). Chemoreceptor gene loss and acquisition via horizontal gene transfer in Escherichia coli. J. Bacteriol. *195*, 3596–3602.

Brown, C.T., Sharon, I., Thomas, B.C., Castelle, C.J., Morowitz, M.J., and Banfield, J.F. (2013). Genome resolved analysis of a premature infant gut microbial community reveals a Varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life. Microbiome *1*, 30.

Daniel, H., Moghaddas Gholami, A., Berry, D., Desmarchelier, C., Hahne, H., Loh, G., Mondot, S., Lepage, P., Rothballer, M., Walker, A., et al. (2014). High-fat diet alters gut microbiota physiology in mice. ISME J. *8*, 295–308.

Fitzsimons, M.S., Novotny, M., Lo, C.-C., Dichosa, A.E.K., Yee-Greenbaum, J.L., Snook, J.P., Gu, W., Chertkov, O., Davenport, K.W., McMurry, K., et al. (2013). Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. Genome Res. *23*, 878–888.

Frank, D.N., St Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. USA *104*, 13780–13785.

Furukawa, S., Fujita, T., Shimabukuro, M., Iwaki, M., Yamada, Y., Nakajima, Y., Nakayama, O., Makishima, M., Matsuda, M., and Shimomura, I. (2004). Increased oxidative stress in obesity and its impact on metabolic syndrome. J. Clin. Invest. *114*, 1752–1761.

Gevers, D., Vandepoele, K., Simillon, C., and Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. *12*, 148–154.

Gill, S.R., Fouts, D.E., Archer, G.L., Mongodin, E.F., Deboy, R.T., Ravel, J., Paulsen, I.T., Kolonay, J.F., Brinkac, L., Beanan, M., et al. (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. J. Bacteriol. *187*, 2426–2438.

Han, X., Kennan, R.M., Davies, J.K., Reddacliff, L.A., Dhungyel, O.P., Whittington, R.J., Turnbull, L., Whitchurch, C.B., and Rood, J.I. (2008). Twitching motility is essential for virulence in Dichelobacter nodosus. J. Bacteriol. *190*, 3323–3335.

Hansen, E.E., Lozupone, C.A., Rey, F.E., Wu, M., Guruge, J.L., Narra, A., Goodfellow, J., Zaneveld, J.R., McDonald, D.T., Goodrich, J.A., et al. (2011). Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. Proc. Natl. Acad. Sci. USA *108* (1), 4599–4606.

Heikkinen, E., Kallonen, T., Saarinen, L., Sara, R., King, A.J., Mooi, F.R., Soini, J.T., Mertsola, J., and He, Q. (2007). Comparative genomics of Bordetella pertussis reveals progressive gene loss in Finnish strains. PLoS ONE *2*, e904.

Hoffman, L.R., Pope, C.E., Hayden, H.S., Heltshe, S., Levy, R., McNamara, S., Jacobs, M.A., Rohmer, L., Radey, M., Ramsey, B.W., et al. (2014). Escherichia coli dysbiosis correlates with gastrointestinal dysfunction in children with cystic fibrosis. Clin. Infect. Dis. *58*, 396–399.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Iida, N., Dzutsev, A., Stewart, C.A., Smith, L., Bouladoux, N., Weingarten, R.A., Molina, D.A., Salcedo, R., Back, T., Cramer, S., et al. (2013). Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. Science *342*, 967–970.

Kinross, J.M., Darzi, A.W., and Nicholson, J.K. (2011). Gut microbiome-host interactions in health and disease. Genome Med. *3*, 14.

Kraal, L., Abubucker, S., Kota, K., Fischbach, M.A., and Mitreva, M. (2014). The prevalence of species and strains in the human microbiome: a resource for experimental efforts. PLoS ONE *9*, e97279.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

Langsjoen, P.H., and Langsjoen, A.M. (2014). Comparison study of plasma coenzyme Q 10 levels in healthy subjects supplemented with ubiquinol versus ubiquinone. Clin. Pharmacol. Drug Dev. *3*, 13–17.

Larsen, N., Vogensen, F.K., van den Berg, F.W.J., Nielsen, D.S., Andreasen, A.S., Pedersen, B.K., Al-Soud, W.A., Sørensen, S.J., Hansen, L.H., and Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS ONE *5*, e9085.

Lee, M.-C., and Marx, C.J. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet. *8*, e1002651.

Levy, R., and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc. Natl. Acad. Sci. USA *110*, 12804–12809.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Morowitz, M.J., Denef, V.J., Costello, E.K., Thomas, B.C., Poroyko, V., Relman, D.A., and Banfield, J.F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc. Natl. Acad. Sci. USA *108*, 1128–1133.

Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., González, A., Fontana, L., Henrissat, B., Knight, R., and Gordon, J.I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science *332*, 970–974.

Neville, B.A., Sheridan, P.O., Harris, H.M.B., Coughlan, S., Flint, H.J., Duncan, S.H., Jeffery, I.B., Claesson, M.J., Ross, R.P., Scott, K.P., and O'Toole, P.W. (2013). Pro-inflammatory flagellin proteins of prevalent motile commensal bacteria are variably abundant in the intestinal microbiome of elderly humans. PLoS ONE *8*, e68919.

Pimenta, A.L., Racher, K., Jamieson, L., Blight, M.A., and Holland, I.B. (2005). Mutations in HlyD, part of the type 1 translocator for hemolysin secretion, affect the folding of the secreted toxin. J. Bacteriol. *187*, 7471–7480.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

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

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. Proc. Natl. Acad. Sci. USA *97*, 14668–14673.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., et al. (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45–50.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., and Alm, E.J. (2012). Population genomics of early events in the ecological differentiation of bacteria. Science *336*, 48–51.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. *23*, 111–120.

Siezen, R.J., Tzeneva, V.A., Castioni, A., Wels, M., Phan, H.T.K., Rademaker, J.L.W., Starrenburg, M.J.C., Kleerebezem, M., Molenaar, D., and van Hylckama Vlieg, J.E.T. (2010). Phenotypic and genomic diversity of Lactobacillus plantarum strains isolated from various environmental niches. Environ. Microbiol. *12*, 758–773.

Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. Nature *480*, 241–244.

Sohet, F.M., Neyrinck, A.M., Pachikian, B.D., de Backer, F.C., Bindels, L.B., Niklowitz, P., Menke, T., Cani, P.D., and Delzenne, N.M. (2009). Coenzyme Q10 supplementation lowers hepatic oxidative stress and inflammation associated with diet-induced obesity in mice. Biochem. Pharmacol. *78*, 1391–1400.

Solheim, M., Aakra, A., Snipen, L.G., Brede, D.A., and Nes, I.F. (2009). Comparative genomics of Enterococcus faecalis from healthy Norwegian infants. BMC Genomics *10*, 194.

Sonti, R.V., and Roth, J.R. (1989). Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics *123*, 19–28.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Vijay-Kumar, M., Aitken, J.D., Carvalho, F.A., Cullender, T.C., Mwangi, S., Srinivasan, S., Sitaraman, S.V., Knight, R., Ley, R.E., and Gewirtz, A.T. (2010). Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science *328*, 228–231.

Zaneveld, J.R., Lozupone, C., Gordon, J.I., and Knight, R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res. *38*, 3869–3879.

Zunino, P., Piccini, C., and Legnani-Fajardo, C. (1994). Flagellate and non-flagellate Proteus mirabilis in the development of experimental urinary tract infection. Microb. Pathog. *16*, 379–385.