Cryptic functional variation in the human gut microbiome

Patrick H. Bradley[1], Katherine S. Pollard[1,2,3,4*]

June 1, 2016

1. Gladstone Institutes @ UCSF, San Francisco, CA.

2. Dept. of Epidemiology & Biostatistics, Division of Biostatistics, UCSF, San Francisco, CA.

3. Institute for Human Genetics, UCSF, San Francisco, CA.

4. Center for Bioinformatics & Molecular Biostatistics, UCSF, San Francisco, CA.

* Corresponding author.

E-mails: patrick.bradley@gladstone.ucsf.edu, katherine.pollard@gladstone.ucsf.edu

Abstract

Background: The human gut microbiome harbors microbes that perform diverse biochemical functions. Previous work suggested that functional variation between gut microbiota is small relative to taxonomic variation. However, these conclusions were largely based on broad pathways and qualitative patterns. Identifying microbial genes with highly variable or invariable abundance across hosts requires a new statistical test.

Results: We develop a model for microbiome gene abundance that allows for differences in means between studies and accounts for the mean-variance relationship in shotgun data. Applying a test based on this model to stool metagenomes from three populations of healthy adults, we discover many significantly variable genes, including components of central carbon metabolism and other pathways comprised primarily of more stable genes. By integrating taxonomic profiles into our test for gene variability, we reveal that Proteobacteria are a major source of variable genes. Stable genes tend to have broad phylogenetic distributions, but several two-component signaling pathways and carbohydrate utilization gene families have relatively constant levels across hosts despite being taxonomically restricted.

Conclusions: Gene-level tests shed light on adaptation to the gut environment, and highlight microbially-encoded functions that may respond to or cause variability in host traits.

Keywords

human gut microbiome, variance, shotgun metagenomics, statistical methods, functional redundancy

1 Background

The microbes that inhabit the human gut encode a wealth of proteins that contribute to a broad range of biological functions, from modulating the human immune system [1, 2, 3] to participating in metabolism [4, 5]. Shotgun metagenomics is revolutionizing our ability to identify and quantify protein-coding genes from these microbes. However, we still lack a comprehensive understanding of the factors that govern host-microbe interactions and microbial fitness within the gut. For example, some bacteria become long-term gut residents while others (e.g., some probiotics [6]) only inhabit the gastro-intestinal (GI) tract transiently. Differences in microbiome composition and diversity have been associated with many diseases. Establishing causality and designing microbiome therapies will require much deeper understanding of microbially-encoded genes and their relationships to human genetics

and the gut environment. This knowledge could not only offer mechanistic explanations for host-microbe associations, but may ultimately help us to engineer the gut microbiome for human health.

Recent screens have shed light on bacterial genes allowing human colonization, such as sugar or polysaccharide utilization genes (e.g. [7, 8]). These screens, while highly informative, are labor-intensive and limited to specific bacterial clades (e.g., *Bacteroides* or *Lactobacilli*). They also require not only that a particular species be culturable, but also that techniques for forward genetics (e.g., transposon insertion) have been developed in that species. Furthermore, it is likely that multiple niches exist within the human gut, and that the metabolic environment of the gut may be influenced by factors like diet [9, 10], variation in immune function [11, 12], prior antibiotic use [13, 14], and host genetics [15]. This suggests that many different symbiotic gut microbial lifestyles are possible and that selective pressures are likely to differ between individual hosts. For these reasons, a computational approach incorporating high-throughput microbiome sequencing data from multiple human populations could offer unique insights.

Gene families that are necessary for life in the gut are expected to have consistent levels across different human hosts, spanning geography and other variables like age and sex. Conversely, gene families that contribute to survival in one particular type of gut environment (e.g., associated with a particular diet) should vary between subjects. Therefore, we propose that the variability of gene families, rather than only their average abundance across sampled metagenomes, is a statistic that can be informative regarding selection in the gut and differences in host-microbe interactions across people.

There is substantial interest in characterizing the extent of functional variation across gut microbiomes [16, 17]. Previous work suggested that the overall variability of gene functions across metagenomes is lower than 1. the variability of those same functions across fully-sequenced genomes [17] and 2. the variability of taxa in the same metagenomes [16] (the "functional stability" hypothesis). However, these analyses tended to be qualitative or binary in nature, as opposed to probabilistic, and/or have been conducted on relatively broad units of function, such as biological pathways or the entire set of annotated functions. On the other hand, microbe-host interactions usually depend upon specific genes, such as colitis-inducing cytolethal distending toxins of *Helicobacter hepaticus* [18] and the enzymes of commensal bacteria that protect against these toxins by producing anti-inflammatory polysaccharide A [19]. Similarly, differences in drug efficacy [20] or side-effects [21] across patients can be attributed to levels of specific microbiome proteins.

To enable high-resolution, quantitative analysis of functional stability in the microbiome, we developed a statistical test that identifies individual gene families whose abundances

are either significantly variable or invariable across samples, without needing to choose a comparison group from either metagenomes or whole genomes. Our method fits gene family abundances to a model and calculates the variance of the residuals, then compares this residual variance statistic to a data-driven null distribution based on the negative binomial distribution. Using simulated data, we show that under certain assumptions, this test has high power (>90%) and controls the false-positive rate appropriately. When these assumptions are violated, we can use simulations to control the false discovery rate (FDR) empirically.

We apply this test to healthy gut metagenomes (n = 123) spanning three different shotgun sequencing studies and find both significantly invariable (3,768) and variable (1,219) gene families (FDR<5%). Many pathways, including some commonly viewed as "housekeeping" or previously identified as stable across gut microbiota (e.g., central carbon metabolism), include significantly variable gene families. Phylogenetic distribution (PD) correlates overall with variability in gene family abundance, and exceptions to this trend highlight functions that may be involved in adaptation, such as two-component signaling and specialized secretion systems. Proteobacteria emerge as a source for genes with the greatest variability in abundance across hosts, suggesting a relationship between inflammation and gene-level differences in what gut microbiota are doing.

2 Results

2.1 The residual variance statistic captures the variability of gene families across hosts

In metagenomics data, different gene families vary widely in average abundance. Gene family abundances can also vary by study, both because of biological differences between populations, and for technical reasons including library preparation, amplification protocol, and sequencing technology. To account for such effects, we fit a linear model of log abundance $D_{g,s}$ for gene g in sample s as a function of a per-gene-family mean G_g, "offsets" $X_{g,y}$ for each study y, and residual variation $\epsilon_{g,s}$:

$$D_{g,s} = G_g + \sum_{y \in Y} I_{y,s} X_{g,y} + \epsilon_{g,s} \tag{1}$$

where $I_{y,s}$ is an indicator variable that is 1 if sample s belongs to study y and 0 otherwise. We take the variance of the residuals, $V_s(\epsilon_{g,s})$, or V_g^{ϵ} for short, as our statistic. This statistic captures how much variance remains after accounting for dataset-specific and gene-family-

specific shifts in abundance (Figure 1).

2.2 A null distribution based on the negative binomial allows detection of significant (in)variability

While calculating the residual variance statistic V_g^ϵ allows a ranking of gene families by variability, by itself, it does not tell us whether this variability (or lack thereof) is surprising. Since there is no straightforward formula for the p-value associated with this statistic, we developed a method to assess significance (Supplemental Figure S1). We choose the null hypothesis that the metagenomic read count data (before any normalization, e.g., for gene length and average genome size) are distributed negative-binomially. The negative binomial distribution is frequently used to model high-throughput sequencing data [22], and can be conceptualized as an overdispersed Poisson where the variance can exceed the mean.

To obtain our null distribution, we allow the mean value to change for different gene families as observed, but fix the overdispersion parameter k that controls the mean-variance relationship across genes. This has similarities to previous approaches to model RNAseq distributions [23] and to identify (in)variable genes from single-cell RNAseq data [24] (see also Discussion). Intuitively, this null hypothesis means that the gene families we identify as significantly variable or invariable will be those whose abundances are over- or under-dispersed, respectively, relative to the median gene family. This choice of null is important: if we instead simply test for high variance, regardless of mean abundance, highly abundant gene families (e.g., single-copy proteins in the bacterial ribosome) are significantly variable despite being nearly universally present at equal abundance in each bacterial genome, because genes with high mean abundance will have high variance in any sequencing experiment. Conversely, thousands of lower-abundance gene families would appear to be significantly invariable simply by virtue of having relatively low read counts.

Based on simulations with similar properties to real data (see Methods, Supplemental Figure S3), we find that this test has high power (> 0.9). The false positive rate (i.e., type I error α) is well-controlled as long as the overdispersion parameter k used in the null distribution is accurately estimated. This appears to be easiest to achieve when fewer than 50% of gene families are significantly variable or invariable. To make the test more robust to estimation of k, we developed a simulation based method to empirically identify significance thresholds that control FDR for both variable and invariable families (Supplemental Table S7).

2.3 Thousands of gene families in the gut microbiome are significantly (in)variable

To describe variation within healthy gut microbiota across different human populations, we randomly selected 123 metagenomes of healthy individuals from the Human Microbiome Project (HMP) [16], controls in a study of type II diabetes (T2D) [25], and controls in a study of glucose control (GC) [26]. These span American, Chinese, and European populations, respectively (see Methods). We map these metagenomes to KEGG Orthology families with ShotMAP [27] and quantify gene family abundance using log-transformed reads per kilobase of genome equivalents (RPKG) [28]. This produces an $n \times m$ data matrix D of log-RPKG abundances consisting of n = 17,417 gene families across m = 123 healthy human gut samples.

Before testing for significantly variable and invariable KEGG families, we conducted a simulation that mirrored this dataset (n = 120, variable-to-invariable gene family ratio between 1:2 and 1:3). The observed α was somewhat higher than the targeted level for the variable gene families (Supplemental Figure S4). We therefore used the simulation results to empirically identify FDR controlling significance thresholds.

We find 2,357 gene families with more variability than expected and 5,432 with less (leaving 9,628 non-significant) at an empirical FDR of 5% (Supplemental Figure S5). Restricting ourselves further to gene families with at least one annotated representative from a bacterial or archaeal genome in KEGG, we obtain 1,219 significantly variable and 3,813 significantly invariable gene families (and 2,194 non-significant). The differences in the residual variation of these gene families can be visualized using a heatmap of the residual $\epsilon_{g,s}$ values (Supplemental Figures S6, S7).

Importantly, the magnitude of the residual variance statistic V_g^ϵ is not the sole determinant of significance, as observed by the overlap in distributions of V_g^ϵ between the variable, invariable, and non-significant gene families. For example, both low-abundance gene families with many zero values and high-abundance but invariable gene families will tend to have low residual variance, but the evidence for invariability is much stronger for the second group. Our test accurately discriminates between these scenarios, tending to call the second group significantly invariable and not the first (Supplemental Figure S5, inset).

2.4 Biological pathways, including those in central metabolism, contain a range of stable and variable components

It has been observed that person-to-person differences in the taxonomic composition of healthy gut microbiomes are much larger than differences in functional composition [16].

This has been interpreted as evidence that diverse gut microbiota are doing the same things [16, 17].

Our results support but also qualify this conclusion. Many of the specific pathways identified as stable (e.g., aminoacyl-tRNA metabolism, central carbon metabolism) have more stable than variable genes in our analysis. However, these pathways also include significantly variable genes (Figure 2). For example, even the highly conserved KEGG module set "aminoacyl tRNA" includes one variable gene at an empirical FDR of 5%, SepRS. SepRS is an O-phosphoseryl-tRNA synthetase, which is an alternative route to biosynthesis of cysteinyl-tRNA in methanogenic archaea [29]. Methanogen abundance has previously been noted to be variable between individual human guts, though this may be due to variability in DNA extraction for archaea and not due to true differences in abundance [30]. Another gene in this category is variable at a weaker level of significance (10% empirical FDR): PoxA, a variant lysyl-tRNA synthetase. Recent experimental work has shown that this protein has a diverged, novel functionality, lysinylating the elongation factor EF-P [31, 32].

By comparison, in the KEGG module set "central carbohydrate metabolism", 77% of the tested prokaryotic gene families were significantly stable, and 5.6% (5 genes) were significantly variable (Figure 2) at an empirical FDR of 5%. In this case, the variable gene families highlight the complexities of microbial carbon utilization. Glucose can be metabolized by two alternative pathways: the more famous Embden-Meyerhof-Parnas (EMP) pathway (i.e., classical "glycolysis"), or the Entner-Doudoroff pathway (ED). Both take glucose to pyruvate, but with differing yields of ATP and electron carriers; ED also allows growth on sugar acids like gluconate [33]. Indeed, while all genes in the "core module" of glycolysis dealing with 3-carbon compounds were significantly invariable across individuals, the ED-specific gene family edd, which takes 6-phosphogluconate to 2-keto-3-deoxy-phosphogluconate (KDPG), was significantly variable according to our test.

We discovered significant variability in abundance for other unusual glycolytic enzymes and enzymes in the tricarboxylic acid cycle (TCA). Multifunctional (K16306, K01622) variants of fructose-bisphosphate aldolase were significantly variable, while the typical FBA enzyme (FbaA) was significantly stable. A subunit of fumarate reductase, frdD, was also significantly variable. Fumarate reductase catalyzes the reverse reaction from the typical TCA cycle enzyme succinate dehydrogenase and can be used for redox balance during anaerobic growth [34]. Conversely, the standard succinate dehydrogenase genes sdhA, sdhB and sdhC were significantly invariable. These results suggest that using our test to identify variable genes within otherwise stable pathways can reveal diverged functionality as well as families that play domain or clade-specific roles.

We also find that the majority of significantly variable gene families annotated to "bacte-

rial secretion system" (16 out of 18) are involved in specialized secretion systems, especially the type III and type VI systems (Figure 2D). These secretion systems are predominantly found in Gram negative bacteria and are often involved in specialized cell-to-cell interactions, between microbes and between pathogens or symbionts and the host. They allow the injection of effector proteins, including virulence factors, directly into target cells [35, 36]. Type VI secretion systems have also been shown to be determinants of antagonistic interactions between bacteria in the gut microbiome [37, 38]. In contrast, gene families in the Sec (general secretion) and Tat (twin-arginine translocation) pathways were nearly all significantly *stable* at an empirical FDR of 5%, with only one gene in each being found to be significantly variable. This contradicts previous suggestions that the Sec and Tat pathways were some of the most variable in the human microbiome [16]. This discrepancy is probably due to our accounting for the mean-variance relationship in shotgun data; the Sec and Tat systems are abundant and phylogenetically diverse [39].

2.5 Phylogenetic distribution correlates with, but does not totally explain, gene family variability

To explore the relationship between gene family taxonomic distribution and variability in abundance across hosts, we constructed trees of the sequences in each KEGG family using ClustalOmega and FastTree. We then calculated phylogenetic distribution (PD), using tree density to correct for the overall rate of evolution [40] (Figure 4a).

The 2,046 stable families with below-median PD were enriched for the pathways "two-component signaling" (FDR-corrected p-value $q = 1.5{\times}10^{-15}$), "starch and sucrose metabolism" ($q = 1.8 \times 10^{-3}$), "amino sugar and nucleotide sugar metabolism" ($q = 0.063$), "ABC transporters" ($q = 2.4 \times 10^{-5}$), and "glycosaminoglycan [GAG] degradation" ($q = 0.053$), among others (Supplemental Table S1). Enriched modules included a two-component system involved in sporulation control ($q = 0.018$), as well as transporters for rhamnose ($q = 0.14$), cellobiose ($q = 0.14$), and alpha- and beta-glucosides ($q = 0.14$ and $q = 0.19$, respectively). These results are consistent with the hypothesis that one function of the gut microbiome is to encode carbohydrate-utilization enzymes the host lacks [41]. Additionally, recent experiments have also shown that the major gut commensal *Bacteroides thetaiotaomicron* contain enzymes adapted to the degradation of sulfated glycans including GAGs [42, 43], and that many *Bacteroides* species can in fact use the GAG chondroitin sulfate as a sole carbon source [44].

Out of the 298 significantly-variable gene families with above-median PD, we found no pathway enrichments but three module enrichments. Only the archaeal ($q = 1.5 \times 10^{-3}$)

and eukaryotic ($q = 8.7 \times 10^{-9}$) ribosomes were enriched, as expected from the abundance patterns of ribosomal proteins from different domains of life across hosts (2b). We also discovered an enrichment for the type VI secretion system ($q = 0.039$). Finally, gene families described as "hypothetical" were also enriched in the variable/high-PD gene set ($p = 2.4 \times 10^{-8}$, odds ratio = 2.2) and depleted in the invariable/low-PD set ($p = 5.4 \times 10^{-13}$, odds ratio = 0.41). Intriguingly, specialized secretion systems were also observed to vary within gut-microbiome-associated species in a strain-specific manner, using a wholly separate set of data [45].

Transporters were also recently observed to show strain-specific variation in copy number across different human gut microbiomes [45], and analyses by Turnbaugh et al. identified membrane transporters as enriched in the "variable" set of functions in the microbiome [17]. However, we mainly find transporters enriched amongst gene families with similar abundance across hosts, despite being phylogenetically restricted (invariable/low-PD genes). Part of this difference is likely due to our stratifying by phylogenetic distribution, which previous studies did not do.

2.6 Proteobacteria are a major source of variable genes

To explore which taxa contribute variable and stable genes, we first computed correlations between phylum relative abundances (predicted using MetaPhlAn2 [46]) and gene family abundances. This analysis revealed that Proteobacterial levels are correlated with abundance of many variable genes (Figure 5b). Proteobacteria are a comparatively minor component of these metagenomes (median = 1%), compared to Bacteroidetes (median = 59%) and Firmicutes (median = 33%). However, some hosts had up to 41% Proteobacteria. Overgrowth of Proteobacteria has been associated with metabolic syndrome [47] and inflammatory bowel disease [48]. Also, Proteobacteria can be selected (over Bacteroidetes and Firmicutes) by intestinal inflammation as tested by TLR5-knockout mice [49], and some Proteobacteria can induce colitis in this background [50], potentially leading to a feedback loop. Thus, the variable gene families we discovered could be biomarkers for dysbiosis and inflammation in otherwise healthy hosts.

We also examined correlations between gene abundance and three taxonomic summary statistics that have been previously linked to microbiome function: average genome size (AGS) [28], the Bacteroidetes/Firmicutes ratio [17, 51], and α-diversity (Shannon index). All of these statistics were *less* often correlated with variable gene families than with invariable or non-significant gene families (see Supplemental Information, Supplemental Figure S9). These statistics therefore do not explain the variability of gene families in this dataset.

2.7 Each major bacterial phylum has a largely unique complement of variable gene families

The variable gene families we identified seem to include both genes whose variance is explained by phylum-level variation (e.g., Proteobacteria), and genes that vary within fine-grained taxonomic classifications, such as strains within species. Also, some gene families may confer adaptive advantages in the gut only within certain taxa. We were therefore motivated to detect lineage-specific variable and invariable gene families, independent of phylum-level trends. To do so, we repeated the test, but using only reads that mapped best to sequences from each of the four most abundant bacterial phyla (Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria). Because we do not necessarily expect the assumption that fewer than 50% of gene families will be significantly (in)variable to hold within each individual phylum, we estimate the average level of overdispersion from the full dataset instead.

Most (77%) gene families showed phylum-specific effects. Invariable gene families tended to agree, but the reverse was true for variable gene families: 19.4% of gene families that were invariable in one phylum were invariable in all, compared to just 0.34% (8 genes) in the variable set. Gene families invariable in all four phyla were enriched for basal cellular machinery, as expected (Supplemental Table S3).

Mirroring results we obtained in Figure 5, Proteobacteria-specific variable gene families also tended to be variable overall (59%); the opposite trend was true for Bacteroidetes- (12%), Firmicutes- (29%), and Actinobacteria-specific (18%) variable gene families (Figure 6A). This supports the hypothesis that Proteobacterial abundance is a dominant driver of functional variability in the human gut microbiome. It further suggests that many overall-variable gene families are not merely markers for a phylum that varies itself (i.e., Proteobacteria), but are also variable at finer taxonomic levels, such as the species or even the strain level [45, 52].

Comparing the two dominant phyla in the gut, Bacteroidetes and Firmicutes, we further observe that the overall proportions of variable and invariable families are similar across pathways, with some exceptions: for example, lipopolysaccharide (LPS) biosynthesis has many invariable gene families in Bacteroidetes and very few in Firmicutes, which we expect given that LPS is primarily made by Gram-negative bacteria. Conversely, both two-component signaling and the PTS system have many more invariable gene families in Firmicutes than in Bacteroidetes (Figure 6B). However, phylum-specific variable gene families tend not to overlap (median overlap: 0%, compared to 46% for invariable gene families). This is even true for pathways where the overall proportion of variable and invariable gene families is similar, such as cofactor and vitamin biosynthesis and central carbohydrate metabolism (Figure 6C).

Furthermore, the enriched biological functions of the phylum-specific variable gene families differ by phylum (Supplemental Table S4). For instance, Proteobacterial-specific variable gene families are enriched (Fisher's test enrichment $q = 0.13$) for the biosynthesis of siderophore group nonribosomal peptides; iron scavenging is known to be important in the establishment of both pathogens (e.g. *Yersinia*) and commensals (*E. coli*) [53]. Another phylum-specific variable function appears to be the Type IV secretion system (T4SS) within Firmicutes ($q = 0.021$): homologs of this specialized secretion system have been shown to be involved in a wide array of biochemical interactions, including the conjugative transfer of plasmids (e.g. antibiotic-resistance cassettes) between bacteria [54]. We conclude that our approach enables the identification of substantial variation within all four major bacterial phyla in the gut, much of which is not apparent when data are analyzed at broader functional resolution or without stratifying by phylum.

2.8 Variable genes are not biomarkers for body mass index, sex or age

To explore associations of gene variability with measured host traits, we used a two-sided partial Kendall's τ test that controls for study effects (Methods). Body mass index, sex, and age were measured in all three studies we analyzed. None of these variables correlated significantly with any variable gene family abundances, even at a 25% false discovery rate. This suggests that major sources of variation in microbiota gene levels, possibly including diet and inflammation, were not measured in these studies.

3 Discussion

Our test is the first to provide a statistical basis for determining stable versus variable functions in the human gut microbiome, a finer level of quantification and specificity than previously possible. Prior work examining how gene families and metabolic pathways vary across healthy human gut microbiome samples has tended to focus either on comparisons to other body sites or environments, or on identifying general or qualitative trends in variability. For example, Segata et al. [55] identified pathways whose abundance differed in the gut versus other body sites. While valuable, this type of analysis answers a distinct question, as it tests the average abundance and not variance, and depends on comparisons between groups. Thus, it would not identify gene families or pathways that these groups might share (e.g., pathways necessary for human symbiosis more broadly). The second type of approach is exemplified by studies from the HMP Consortium [16] and Turnbaugh et al. [17]. Both

of these studies presented data showing that abundances of several Clusters of Orthologous Groups (COGs) and KEGG modules across the metagenomes of human subjects varied less than phylum abundances among the same samples. However, it is not clear *a priori* whether phylum-level taxonomic variation should indeed be comparable to functional variation, since many functions are so widely conserved, and since many of the tested pathways actually include diverse biological functions (e.g., carbon metabolism). Moreover, these analyses do not identify individual gene families that could break this overall trend.

Turnbaugh et al. [17] also compared the variance of gene families in particular pathways across metagenomes to the variance across sequenced whole genomes. They concluded that biological pathways tended to vary less in metagenomes than in whole genomes. However, there are barriers to extending this test to individual gene families or biological pathways. First, a set of representative whole genomes must be chosen to construct a null distribution. Choosing an appropriate set of genomes to compare to a metagenome is not trivial, especially since the most appropriate genomes may not have been sequenced. Second, samples from a mixture of genomes should have lower variance than samples from individual genomes in that mixture; accounting for this bias could be difficult as it would depend on the precise composition of the sample. This study did classify gene families into "stable" versus "variable" subsets, by looking for families that were observed in all samples versus not observed in at least one. However, while this approach is useful and intuitively appealing, it is more descriptive or qualitative than our method, and would miss gene families that were reliably observed, but whose abundances still varied more than expected between subjects.

We find that basic microbial cellular machinery, such as the ribosome, tRNA-charging, and primary metabolism, are universal functional components of the microbiome, both in general and when each individual phylum is considered separately. This finding is consistent with previous results [17], and indeed, is not surprising given the broad conservation of these processes across the tree of life. However, we also identify candidate invariable gene families that have narrower phylogenetic distributions. These include, for example, proteins involved in two-component signaling, starch metabolism, and glycosaminoglycan metabolism. Previous experimental work has underscored the importance of some of these pathways in gut symbionts: for instance, multiple gut-associated Bacteroides species are capable of using the glycosaminoglycan chondroitin sulfate as a sole carbon source [42], and the metabolism of resistant starch in general is thought to be a critical function of the human microbiome. These results suggest that the method we present is capable of identifying protein-coding gene families that contribute to fitness of symbionts within the gut.

We also identify significantly variable gene families, including specialized secretion systems, e.g., the T6SS. Phylum-specific tests also reveal that gene families involved in the T6SS

are also variable even within Proteobacteria. We find that variable but broadly-conserved gene families include many genes of unknown function, and that these gene families tend to correlate with Proteobacterial abundance (though others also correlate with the abundance of Firmicutes, Bacteroidetes, and Actinobacteria). While fewer in number, we also find invariable gene families whose function is not yet annotated; these gene families may represent functions that are either essential or provide advantages for life in the gut, and may therefore be particularly interesting targets for experimental follow-up (e.g., assessing whether strains in which these gene families have been knocked out in fact have slower growth rates, either in vitro or in the gut).

Though the interpretation of invariable gene families is potentially more straightforward, variable gene families have a variety of ecological interpretations, e.g., first-mover effects, drift, host demography, and selection within particular gut environments. Computationally distinguishing among these possibilities is likely to present challenges. For example, distinguishing selection from random drift will probably require longitudinal data and appropriate models. Separating effects of host geography, genetics, medical history, and lifestyle will be possible only when richer phenotypic data is available from a more diverse set of human populations. To control for study bias and batch effects, it will be important to include multiple sampling sites within each study.

While statistical tests focused on differences in variances are not yet common throughout genomics, there is some recent precedent using this type of test to quantify the gene-level heterogeneity in single-cell RNA sequencing data [56, 24], and to identify variance effects in genetic association data [57]. Like Vallejos et al. [24], we model gene counts using the negative binomial distribution, and identify both significantly variable and invariable genes, although we frame our method as a frequentist hypothesis test as opposed to a Bayesian hierarchical model. Unlike previous approaches in this domain, the method we describe does not explicitly decompose biological from technical noise, and therefore does not require the use of experimentally-spiked-in controls, which are not present in most experiments involving sequencing of the gut microbiome.

A similar statistical method for detecting significant (in)variability such as the one we present here could also be applied to other biomolecules measured in counts, such as metabolites, proteins, or transcripts. Performing such analyses on human microbiota would reveal patterns in the variability in the usage of particular genes, reactions, and pathways, which would expand on our investigation of potential usage based on presence in the DNA of organisms in host stool. Another important extension is to generalize our method for comparing hosts from different pre-defined groups (disease states, countries, diets) to identify gene families that are stable in one group (e.g., healthy controls) but variable in another

(e.g., patients). Since metagenomic samples contain substantial heterogeneity, investigating group differences in functional variability could allow the detection of different trends from the more common comparison of means.

4 Conclusion

This study present a novel statistical method that provides a finer resolution estimate of "functional redundancy" [58] in the human microbiome than was previously possible. We found that most biological pathways, including tRNA charging, central carbon metabolism, and bacterial secretion, have both invariable and variable components. Some variable genes have surprisingly broad phylogenetic distributions, and Proteobacteria emerge as a major source of variable genes. Since Proteobacteria have also been linked to inflammation and metabolic syndrome [47], we speculate that baseline inflammation may be one variable influencing functions in the gut microbiome.

5 Methods

5.1 Data collection and processing

Stool metagenomes from healthy human guts were obtained from three sources:

1. two American cohorts from the Human Microbiome Project [16], n = 42 samples selected;

2. a Chinese cohort from a case-control study of type II diabetes (T2D) [25], n = 44 samples from controls with neither type II diabetes nor impaired glucose tolerance;

3. and a European cohort from a case-control study of glucose control [26], n = 37 samples from controls with normal glucose tolerance.

After downloading these samples from NCBI's short read archive (SRA), the FASTA-formatted files were mapped to KEGG Orthology (KO) [59] protein families as previously described [27]. For consistency, each sample was rarefied to a depth of 1.5×10^7 reads, and additionally, as reads from HMP were particularly variable in length, they were therefore trimmed to a uniform length of 90 bp.

For each sample, we used ShotMAP to detect how many times a particular gene family matched a read ("counts"; we add one pseudocount for reasons described below). The bit-score cutoff for matching a protein family was selected based on the average read length of

each sample as recommended previously [27]. For every gene family in every sample, we also computed the average family length (AFL), or the average length of the matched genes within a family. Finally, we also computed per-sample average genome size using MicrobeCensus [28] (`http://github.com/snayfach/MicrobeCensus`). These quantities were used to estimate abundance values in units of RPKG, or reads per kilobase of genome equivalents [28].

These RPKG abundance values were strictly positive with a long right tail and highly correlated with the variances (Spearman's r = 0.99). This strong mean-variance relationship is likely simply because these abundances are derived from counts that are either Poisson or negative-binomially distributed. We therefore took the natural log of the RPKG values as a variance stabilizing transformation. Because $log(0)$ is infinite, we add a pseudocount before normalizing the counts and taking the log transform. Since there is no average family length (AFL) when there are no reads for a given gene family in a given sample, we impute it in those cases using the average AFL across samples.

5.2 Model fitting

We fit a linear model to the data matrix of log-RPKG D of log-RPKG described above, with n gene-families by m samples, to capture gene-specific and dataset-specific effects:

$$D_{g,s} = G_g + \sum_{y \in Y} I_{y,s} X_{g,y} + \epsilon_{g,s} \tag{2}$$

where $g \in [1, n]$ is a particular gene family, $s \in [1, m]$ is a particular sample, G_g is the grande or overall mean of log-RPKG $\frac{\sum_s D_{g,s}}{m}$ for a given gene family g, Y is the set of studies, $I_{y,s}$ is an indicator variable valued 1 if sample s is in study y and 0 otherwise, $X_{g,y}$ is a mean offset for gene family g in study y, and the residual for a given gene family and sample are given by $\epsilon_{g,s}$. For each gene family, the variance across samples of these $\epsilon_{g,s}$, which we term the "residual variance" or V_g^ϵ, becomes our statistic of interest.

Overall trends in these data are explained well by this model, with an $R^2 = 0.20$. The residuals, which are approximately symmetrically distributed around 0, represent variation in gene abundance not due to study effects.

5.3 Modeling residual variances under the null distribution

Having calculated this statistic V_g^ϵ for each gene family g, we then need to compare this statistic to its distribution under a null hypothesis H_0. This requires us to model what the data would look like if in fact there were no surprisingly variable or invariable gene families. To do this, we use the negative binomial distribution to model the original count data (before

adding pseudocounts and normalization to obtain RPKG).

The negative binomial distribution is commonly used to model count data from high throughput sequencing. It can be conceptualized as a mixture of Poisson distributions with different means, which themselves follow a Gamma distribution. Like the Poisson distribution, the negative binomial distribution has an intrinsic mean-variance relationship. However, instead of a single mean-variance parameter as in the Poisson, the negative binomial can be described with two, a mean parameter and a "size" parameter, which we refer to here as k such that $k = \frac{\mu^2}{\sigma^2 - \mu}$. k ranges from $(0, \infty)$, with smaller values corresponding to more overdispersion (i.e., higher variance given the mean) and larger values approaching, in the limit, the Poisson distribution.

To model the case where no gene family has unusual variance given its mean value, i.e., our null hypothesis, we assume that the data are negative-binomially distributed with the observed means $\mu_{g,y}$ for each gene g and study y, but where the amount of overdispersion is modeled with a single size parameter k_y for each study y:

$$\begin{aligned} H_0: &\quad V_g^\epsilon = V_g^\epsilon | D_{g,s} \sim NB(\mu_{g,y},\, k_y) \\ H_{alt}: &\quad V_g^\epsilon \neq V_g^\epsilon | D_{g,s} \sim NB(\mu_{g,y},\, k_y) \end{aligned}$$

To estimate this $\widehat{k_y}$, the overall size parameter for a given study y, we estimate the mode of per-gene-family size parameters $k_{g,y}$ within data set y, using the method-of-moments estimator for each $k_{g,y}$. We accomplish this by fitting a Gaussian kernel density estimate to the log-transformed $k_{g,y}$ values, and then finding the $\widehat{k_y}$ value that gives the highest density. (From simulations, we found that the mode method-of-moments was more robust than the median or harmonic mean: see Supplemental Figure S2.) We can then easily generate count data under this null distribution, add a pseudocount and normalize by AFL and AGS, fit the above linear model, and obtain null residual variances $V_g^{\epsilon 0}$ using exactly the same procedure described above.

Statistical significance is then obtained by a two-tailed test:

$$p_g = \frac{\#\left(\left(\frac{V_g^{\epsilon 0} - \overline{V_g^{\epsilon 0}}}{\overline{V_g^{\epsilon 0}}}\right)^2 \geq \left(\left(\frac{V_g^{\epsilon} - \overline{V_g^{\epsilon 0}}}{\overline{V_g^{\epsilon 0}}}\right)^2\right)\right) + 1}{B + 1}$$

Here, B refers to the number of null test statistics $V_g^{\epsilon 0}$ (in this case, $B = 750$), and the overlined test statistics refer to their mean across the null distribution.

The resulting p-values are then corrected for multiple testing by converting to FDR q-values using the procedure of Storey et al. [60] as implemented in the `qvalue` package in R [61]. An alternative approach to determining significance is based on the bootstrap.

While using a parametric null distribution allows us to explicitly model the null hypothesis, it also breaks the structure of covariance between gene families, which may be substantial because genes are organized into operons and individual genomes within a metagenome. This structure can, optionally, be restored using a strategy outlined by Pollard and van der Laan [62]. Instead of using the test statistics $V_g^{\epsilon_0}$ obtained under the parametric null as is, we can use these test statistics to center and scale bootstrap test statistics $V_g^{\epsilon\prime}$, which we derive from applying a cluster bootstrap with replacement from the real data and then fitting the above linear model (2) to the resampled data to obtain bootstrap residual variances:

$$V_g^{\epsilon_0\prime} = \left(\left(\frac{V_g^{\epsilon\prime} - \overline{V_g^{\epsilon\prime}}}{sd(V_g^{\epsilon\prime})} \right) \times sd(V_g^{\epsilon_0}) \right) + \overline{V_g^{\epsilon_0}}$$

A similar non-parametric bootstrap approach has previously been successfully applied to testing for differences in gene expression [63].

As expected, when the residuals are plotted in a heatmap as in Figure S6, variable gene families are generally brighter (i.e., more deviation from the mean) than invariable gene families, though not exclusively: this is because our null distribution, unlike the visualization, models the expected mean-variance relationship. We visualize this information by scaling each gene family by its expected standard deviation under the negative binomial null (i.e., by the mean root variance $\sum_{b\in[1,B]} \sqrt{V_{g_b}^{\epsilon_0}}/B$) (Figure S7).

5.4 Power analysis

The test we present controls α as expected if the correct size parameter k is estimated from the data (Supplemental Figure S2a-b). Estimating this parameter accurately is known to be difficult, however, particularly for highly over-dispersed data [64], and in this case we must also estimate this parameter from a mixture of true positives and nulls. We find that the mode of per-gene-family method-of-moments estimates is more robust to differences in the ratio of variable to invariable true positives (Supplemental Figure S2e-g) than the median or harmonic mean (the harmonic mean mirrors the approach in Yu et al. [23]).

Power analysis was performed on simulated datasets comprising three simulated studies. For each study, 1,000 gene families were simulated over $n \in \{60, 120, 480, 960\}$ samples. Null data were drawn from a negative-binomial distribution with a randomly-selected size parameter k in common to all gene families, which was drawn from a log-normal distribution (log-mean= -0.65, sd= 0.57). Gene family means were also drawn from a log-normal (log mean= 2.94, sd = 2.23). True positives were drawn from a similar negative-binomial distribution, but where the size parameter was multiplied by an effect size z (for variable

gene families) or its reciprocal $1/z$ (for invariable gene families). The above test was then applied to the simulated data, and the percent of Type I and II errors was calculated by comparing to the known gene family labels from the simulation. Using similar parameters to those estimated from our real data, we see that α decreases and power approaches 1 with increasing sample size (see Supplemental Figure S3) and that $n = 120$ appears to be sufficient to achieve control over α.

However, at $n = 120$, we also noted that α appeared to be greater for variable vs. invariable gene families (Supplemental Figure S4), possibly because accurately detecting additional overdispersion in already-overdispersed data may be intrinsically difficult. We therefore performed additional simulations to determine q-value cutoffs corresponding to an empirical FDR of 5%. We calculated appropriate cutoffs based on datasets with 43% true positives and a variable:invariable gene family ratio ranging from 0.1 to 10, taking the median cutoff value across these ratios (Supplementary Table S7). Using these cutoffs, the overall dataset had 45% true positives and a variable:invariable gene family ratio of 0.43.

5.5 Calculating phylogenetic distribution of gene families

The phylogenetic distribution (PD) of KEGG Orthology (KO) families was estimated using tree density [40]. We first obtained sequences of each full-length protein annotated to a particular KO, and then performed a multiple alignment of each family using ClustalOmega [65]. These multiple alignments were used to generate trees via FastTree [66]. For both the alignment and tree-building, we used default parameters for homologous proteins.

For all families represented in at least 5 different archaea and/or bacteria (6,703 families total), we then computed tree densities, or the sum of edge lengths divided by the mean tip height. Using tree density instead of tree height as a measure of PD corrects for the rate of evolution, which can otherwise cause very highly-conserved but slow-evolving families like the ribosome to appear to have a low PD [40]. Empirically, this measure is very similar to the number of protein sequences (Supplemental Figure S8), but is not as sensitive to high or variable rates of within-species duplication: for example, families such as transposons, which exhibit high rates of duplication as well as copy-number variation between species, have a larger number of sequences than even very well-conserved proteins such as RNA polymerase, but have similar or even lower tree densities, indicating that they are not truly more broadly conserved.

Many protein families (8,931 families) did not have enough observations in order to reliably calculate tree density, with almost all of these being annotated in only a single bacterium/archaeum. For these, we predicted their PD by extrapolation. To predict PD,

we used a linear model that predicted tree density based on the total number of annotations (including annotations in eukaryotes). In five-fold cross-validation, this model actually had a relatively small mean absolute percentage error (MAPE) of 13.1%. We also considered a model that took into account the taxonomic level (e.g., phylum) of the last common ancestor of all organisms in which a given protein family was annotated, but this model performed essentially identically (MAPE of 13.0%). Predicted tree densities are given in Supplemental Table S6. The PD of gene families varied from 1.2 (an iron-chelate-transporting ATPase only annotated in *H. pylori*) to 434.9 (the rpoE family of RNA polymerase sigma factors).

5.6 Gene family enrichment

We were interested in whether particular pathways were enriched in several of the gene family sets identified in this work. For subsets of genes (such as those with specifically low PD), a 2-tailed Fisher's exact test (i.e., hypergeometric test) was used instead to look for cases in which the overlap between a given gene set and a KEGG module or pathway was significantly larger or smaller than expected. The background set was taken to be the intersection of the set of gene families observed in the data with the set of gene families that had pathway- or module-level annotations. p-values were converted to q-values as above. Finally, enrichments were enumerated by selecting all modules or pathways below $q \leq 0.25$ that had positive odds-ratios (i.e., enriched instead of depleted).

5.7 Associations with clinical and taxonomic variables

We were interested in using a non-parametric approach to detect association of residual RPKG with clinical and taxonomic variables (e.g., the inferred abundance of a particular phylum via MetaPhlan2). To take into account potential study effects in clinical and taxonomic variables without using a parametric modeling framework, we used partial Kendall's τ correlation as implemented in the ppcor package for R [67], coding the study effects as binary nuisance variables. Kendall's τ was used over Spearman's ρ because of better handling of ties (an issue with taxonomic variables especially, since many, particularly at the finer-grained levels, were often zero). The null distribution was obtained by permuting the clinical/taxonomic variables within each study 250 times, and then re-assessing the partial τ. Finally, p-values were calculated by taking the fraction of null partial correlations equally or more extreme (i.e., distant from zero) than the real partial correlations.

Phylum-level relative abundances were predicted from the shotgun data using MetaPhlAn2 with the `--very-sensitive` flag [46].

5.8 Phylum-specific tests

We created taxonomically-restricted data sets in which the abundance of each gene family was computed using only metagenomic reads aligning best to sequences from each of the four most abundant bacterial phyla (Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria). Phylum-specific data were obtained from the overall data as follows. First, the NCBI taxonomy was parsed to obtain species annotated below each of the four major bacterial phyla (Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria); these species were then matched with KEGG species identifiers. Next, the original RAPSearch2 [68] results were filtered, so that the only reads remaining were those for which their "best hit" in the KEGG database originally came from the genome of a species belonging to the specific phylum in question (e.g., *E. coli* for Proteobacteria). Finally, when performing the test, normalization for average genome size was accomplished by normalizing gene family counts by the median abundance of a set of 29 bacterial single-copy gene families [69], which had been filtered in the same phylum-specific way as all other gene families; this approach is similar to the MUSiCC method for average genome size correction [70]. This also controls for overall changes in phylum abundance. Finally, $\widehat{k_y}$ values for individual studies were estimated based on the non-phylum-restricted data, since the expectation that $< 50\%$ of gene families were differentially variable might not hold for each individual phylum. We used the same q-value cutoffs as in the overall test to set an estimated empirical FDR (Supplementary Table S7). Otherwise, tests were performed as above.

5.9 Codebase

The scripts used to conduct the test and related analyses are available at the following URL:

`http://www.bitbucket.org/pbradz/variance-analyze`

Counts of reads mapped to KEGG Orthology (KO) groups and average family lengths for all of the samples used in this study can be obtained at FigShare:

- `https://figshare.com/s/fcf1abf369155588ae41` (overall)
- `https://figshare.com/s/90d44cffdfb1d214ef83` (phylum-specific)

6 Author contributions

PHB performed the experiments and analyses. PHB and KSP developed the test, designed the experiments, wrote the paper, and read and approved the final manuscript.

7 Declarations

7.1 Acknowledgements

The authors would like to thank Stephen Nayfach for downloading and organizing metagenomic data and metadata, and for providing and checking code for metagenome annotation, Dongying Wu for suggesting the tree density metric to measure phylogenetic distribution, Aram Avila-Herrera for help with phenotype-to-abundance associations, and Clifford Anderson-Bergman, along with other members of the Pollard group, for helpful discussions. Funding for this research was provided by NSF grant DMS-1069303, Gordon & Betty Moore Foundation grant #3300, and institutional funds from the Gladstone Institutes.

7.2 Information about HMP clinical data

Clinical covariates for HMP were obtained from dbGaP accession #phs000228.v3.p1. Funding support for the development of NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A) was provided by the NIH Roadmap for Medical Research. Clinical data for HMP-A were jointly produced by the Baylor College of Medicine and the Washington University School of Medicine. Sequencing data for HMP-A were produced by the Baylor College of Medicine Human Genome Sequencing Center, The Broad Institute, the Genome Center at Washington University, and the J. Craig Venter Institute. These data were submitted by the EMMES Corporation, which serves as the clinical data collection site for the HMP. Authors read and agreed to abide by the Genomic Data User Code of Conduct.

8 Figures

Figure 1: **The residual variance statistic captures variation in gene families after accounting for between-study variation.** The left panels ("original abundances") show filled circles representing log-abundance (RPKG) for gene families from the KEGG Orthology (KO), with per-study means shown in solid horizontal lines and the distance from these means shown as dashed vertical lines. The right hand panels ("residuals") show the same gene families after fitting a linear model that accounts for these per-study means, with an accompanying density plot showing the distribution of these residuals. V_g^ϵ values in bold underneath density plots are the calculated variances of these residuals. These gene families are sets of orthologs corresponding to A) the waaL family of lipopolysaccharide O-antigen ligases, an immunogenic component of Gram-negative bacterial outer membranes, and B) the vicR family of OmpR family transcriptional regulators that is involved in two-component *vic* signaling, is conserved across many Gram-positive bacteria, and is essential in *Streptococcus pneuomoniae* [71]. Despite having similar overall mean log-abundances and similar magnitude study-specific effects, waaL has much higher residual variance across individual metagenomes than vicR.

Figure 2: **Most pathways include a mixture of both variable and invariable gene families.** A) Stacked bar plots show the fraction of invariable (blue), non-significant (gray), and variable (red) gene families annotated to KEGG Orthology pathway sets (rows), at different false discovery rate (FDR) cutoffs (color intensity). Only gene families with at least one annotated bacterial or archaeal homolog are counted. B) Fraction of strongly invariable, non-significant, and strongly variable gene families within the ribosomes of different kingdoms. Row labels with only one kingdom indicate gene families unique to that kingdom, while rows with multiple kingdoms (e.g. "Eukaryotes/archaea") indicate gene families shared between these two kingdoms. As expected, the bacterial ribosome is completely invariable.

Figure 3: **Variable and invariable gene families within broad biological pathways separate by gene function.** A-C) Heatmaps showing scaled residual *log*-RPKG for gene families (rows) involved in A) tRNA metabolism, B) central carbohydrate metabolism, and C) bacterial secretion systems. Variable (red) and invariable (blue) gene families are clustered separately, as are samples within a particular study (columns). *log*-RPKG values are scaled by the expected variance from the negative-binomial null distribution.

Figure 4: **Phylogenetic distribution (PD) of gene families partially explains gene family variability.** Scatter plot shows log_{10} PD (x-axis) vs. log_{10} residual variance statistic (y-axis). Red points are significantly variable while blue points are significantly invariable. Gene families in specific functional groups are also highlighted in different colors, specifically the bacterial ribosome (pale green), the type VI secretion system (or "T6SS"; orange), the KinABCDE-Spo0FA sporulation control two-component signaling (yellow), and hypothetical genes (tan squares). Gene families that are significantly invariable (ribosome and sporulation control) or significantly variable (hypothetical genes and the T6SS) at an estimated 5% FDR are outlined in black. The bacterial ribosome, as expected, has very high PD and is strongly invariable. The Type VI secretion system genes, in contrast, are conserved but variable, while some genes involved in the Kin-Spo sporulation control two-component signaling pathway have low PD but are invariable. Only gene families with at least one annotated bacterial or archaeal homolog are shown.

Figure 5: **Variable gene families correlate with the predicted abundance of Proteobacteria.** Bar plots give the fraction of gene families in each category (significantly invariable, non-significant, and significantly variable, 5% FDR) that are significantly correlated to predicted relative abundances of phyla, as assessed by MetaPhlan2, using partial Kendall's τ to account for study effects and a permutation test to assess significance. Asterisks give the level of significance by chi-squared test of non-random association between gene family category and the number of significant associations. (***: $p \leq 10^{-8}$ by chi-squared test after Bonferroni correction; **: $p \leq 10^{-4}$.)

Figure 6: **Phylum-specific tests reveal hidden variability in the most prevalent bacterial phyla.** A) Bars indicate the fraction of phylum-specific variable gene families that were also variable overall (red, "both tests") or that were specific to a particular phylum (yellow, "phylum-specific test only"). B) For the Bacteroidetes- (left) and Firmicutes- (right) specific tests, the proportion of invariable (blue), non-significant (gray), and variable (red) gene families, at an estimated 5% FDR (using cutoffs from overall test). Pathways with at least 5 total gene families across both phyla are shown. C) Rectangular Venn diagrams showing the proportion of Bacteroides-specific (left), shared (center, bright), and Firmicutes-specific (right) invariable (blue) and variable (red) gene families for each of the pathways enumerated in B.

9 Additional Files

Figure S1: **Schematic shows overview of data processing and method.** A) Data is processed by taking reads from multiple datasets (represented by letters here) with a certain number of samples (represented by S_A, S_B, etc.). These reads will eventually map to multiple gene families G. MicrobeCensus [28] is used to estimate average genome size, while Shotmap [27]is used to map reads, yielding both matrices of counts (right hand side) and matrices of average lengths of the best-hit proteins ("average family length" or AFL). AFL and AGS estimates are used to normalize counts. B) We calculate our statistic and assign p-values as follows. First, we normalize counts from Shotmap using AFL and AGS, log-transform the resulting reads per kilobase of genome (RPKG), then apply a simple linear model to fit dataset- and gene-family-specific effects. The resulting residuals ("residual log RPKG") form a matrix of G genes by S_A+S_B+S_C samples. We take the variance across all samples for each gene to obtain a 1xG vector of residual variances. To get a null distribution, we can either use data generated from a negative binomial fit, or, optionally, from a negative binomial fit integrated with (shaded section) bootstrap resampling. For the negative binomial fit, from the count matrices, we estimate the mean of each gene in each dataset, as well as dataset-specific overdispersion parameters k. We then use these to make simulated count datasets ("$\times$ B" indicating that this card is replicated once for each of B simulations), which we process as in the case of the real data, yielding simulated log-RPKG matrices and simulated residual variances for each gene family. For the resampling (if applicable), we sample with replacement from each count dataset, yielding resampled counts. We process these in the same way to obtain resampled residual variances. Finally, if using the resampled data, we center and scale the resampled residual variances using per-gene-family means and standard deviations from the simulated residual variances; otherwise, we simply take the values from applying the test to the negative binomial simulations. These form the background distribution (bottom panel, solid curve) for each gene in G ("$\times$ G" indicating that this card is replicated once for each of G genes). The actual observed residual variance (dashed line) is then compared to this distribution to obtain p-values (gray shaded area).

Figure S2: **Size parameter estimation affects power and α.** For each mock dataset y, simulated null data is generated from a negative binomial distribution, fixing the size parameter k_y but allowing the mean $\mu_{g,y}$ to vary for each of 1,000 genes; simulated true-positive gene families are drawn from a negative-binomial distribution with size equal to zk_y or k_y/z, where z is the effect size. A-C) The choice of estimator affects the accuracy of size estimates. The mode method-of-moments estimator (C, y-axis) more accurately estimates the true size specified in the simulation (x-axis) than the harmonic mean (A, y-axis) or median (B, y-axis), and is more tolerant to differences in the ratio of true-positive variable and invariable gene families (colors). D-E) When the size parameter is known, α (D) and power (E) are well controlled, with α approximately equal to 0.05 at $p \leq 0.05$ and power approaching 1. Here, each simulation comprises three mock studies with different size parameters, mirroring our actual data. Bar heights are means from 4 simulations and error bars are ± 2 SD. The proportion of variable:invariable gene families was 0.5 and 44% of genes were true positives.

Figure S3: **Size parameter estimation affects power and α.** α (A) is minimized and power (B) is maximized when the mode method-of-moments estimator is used to get estimates of the study-specific dispersion parameters $\widehat{k_y}$. Bars are from 4 simulations. The proportion of variable:invariable gene families was 0.4 and 43% of genes were true positives.

Figure S4: **The mode estimator is robust to changes in the proportion of true positives and the ratio of variable to invariable gene families.** α (A-C) and power (D-F) as a function of the proportion of true positives (x-axis) and the ratio of variable to invariable true positives (y-axis) for $n = 120$. $\alpha = 0.05$ and power $= 1$ are shown in color-bars to the left of each heatmap for reference. α and power are calculated overall (left), for variable gene families (center), and for invariable gene families (right). In general, α is better controlled for the invariable gene families than for the variable gene families; we therefore use different empirical cutoffs for each set of genes.

Figure S5: **We identify significantly variable and invariable gene families.** Density plots of distributions of residual variance (V_G) statistics for significantly invariable (blue dashed line), non-significant (black solid line), and significantly variable (red dashed line) gene families. The distributions have the expected trend (e.g., significantly variable gene families tend to have higher residual variance) but also overlap, indicating the importance of the calculated null distribution. The inset shows the proportion of zero values for the non-significant (black) and significantly invariable (blue) gene families with V_G falling in the lowest range (vertical dashed lines), indicating that the test differentiates between gene families that only appear invariable because they have few observations, and gene families that are consistently abundant yet invariable.

Figure S8: **Number of leaves is correlated with tree density, but tree density corrects for the overall rate of evolution.** The number of leaves (i.e., individual sequences) is plotted vs. tree density on a log-log scatter plot, with each circle representing one gene family. Two outliers with lower density than expected are plotted in colors: a putative transposase (green) and a *Staphylococcus* leukotoxin (red). Both families have large numbers of sequences from the same organism.

Figure S9: **Variable gene families are less-often correlated to measured host characteristics.** A) Bar plots give the fraction of gene families with at least one bacterial or archaeal representative in each category (significantly invariable, non-significant, and significantly variable) that are significantly correlated to various sample characteristics, using partial Kendall's τ to account for study effects and a permutation test to assess significance. These sample characteristics are average genome size (AGS), the ratio of Bacteroidetes to Firmicutes (B/F ratio), and a measure of α-diversity (Shannon index). (***: $p \leq 10^{-8}$ by chi-squared test after Bonferroni correction; **: $p \leq 10^{-4}$.)

Figure S6: **Heatmap showing significantly variable and invariable gene families (unscaled).** Heatmap showing residual *log*-RPKG abundances (i.e., after normalizing for between-study effects and gene-specific abundances) of significantly invariable (blue) and significantly variable (red) gene families. Variable and invariable gene families are clustered separately, while samples are clustered within each dataset.

Figure S7: **Heatmap showing significantly variable and invariable gene families (scaled).** As with S6, but residual *log*-RPKG abundances scaled by their expected variance under the negative binomial null model (see Methods).

Table S1: Module and pathway enrichments for variable and invariable gene sets (Fisher's exact test $q \leq 0.25$).

Table S2: Module and pathway enrichments for variable/high-PD and invariable/low-PD gene sets (Fisher's exact test $q \leq 0.25$).

Table S3: Module and pathway enrichments for gene families with invariable abundances in every phylum-specific test (Fisher's exact test, $q \leq 0.25$).

Table S4: Module and pathway enrichments for gene families variable in each phylum-specific test (Fisher's exact test, $q \leq 0.25$).

Table S5: SRA IDs and characteristics (read length, average genome size from MicrobeCensus) for samples used in this study.

Table S6: Predicted tree densities.

Table S7: q-value cutoffs to reach a given empirical FDR, estimated from simulation.

empirical FDR	q value cutoff, variable	q value cutoff, invariable
5%	0.0238	0.108
10%	0.0669	0.180
25%	0.181	0.294

10 References

References

[1] Slack, E., Hapfelmeier, S., Stecher, B., Velykoredko, Y., Stoel, M., Lawson, M.A.E., Geuking, M.B., Beutler, B., Tedder, T.F., Hardt, W.-D., Bercik, P., Verdu, E.F., McCoy, K.D., Macpherson, A.J.: Innate and adaptive immunity cooperate flexibly to maintain host-microbiota mutualism. Science (New York, N.Y.) **325**(5940), 617–20 (2009). doi:10.1126/science.1172747

[2] Atarashi, K., Tanoue, T., Shima, T., Imaoka, A., Kuwahara, T., Momose, Y., Cheng, G., Yamasaki, S., Saito, T., Ohba, Y., Taniguchi, T., Takeda, K., Hori, S., Ivanov, I.I., Umesaki, Y., Itoh, K., Honda, K.: Induction of colonic regulatory T cells by indigenous Clostridium species. Science (New York, N.Y.) **331**(6015), 337–41 (2011). doi:10.1126/science.1198469

[3] Hapfelmeier, S., Lawson, M.A.E., Slack, E., Kirundi, J.K., Stoel, M., Heikenwalder, M., Cahenzli, J., Velykoredko, Y., Balmer, M.L., Endt, K., Geuking, M.B., Curtiss, R., McCoy, K.D., Macpherson, A.J.: Reversible microbial colonization of germ-free mice reveals the dynamics of IgA immune responses. Science (New York, N.Y.) **328**(5986), 1705–9 (2010). doi:10.1126/science.1188454

[4] Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J., Buhler, J.D., Gordon, J.I.: Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science (New York, N.Y.) **307**(5717), 1955–9 (2005). doi:10.1126/science.1109051

[5] Wikoff, W.R., Anfora, A.T., Liu, J., Schultz, P.G., Lesley, S.A., Peters, E.C., Siuzdak, G.: Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. Proceedings of the National Academy of Sciences of the United States of America **106**(10), 3698–703 (2009). doi:10.1073/pnas.0812874106

[6] McNulty, N.P., Yatsunenko, T., Hsiao, A., Faith, J.J., Muegge, B.D., Goodman, A.L., Henrissat, B., Oozeer, R., Cools-Portier, S., Gobert, G., Chervaux, C., Knights, D., Lozupone, C.A., Knight, R., Duncan, A.E., Bain, J.R., Muehlbauer, M.J., Newgard, C.B., Heath, A.C., Gordon, J.I.: The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. Science translational medicine **3**(106), 106–106 (2011). doi:10.1126/scitranslmed.3002701

[7] Licandro-Seraut, H., Scornec, H., Pédron, T., Cavin, J.-F., Sansonetti, P.J.: Functional genomics of Lactobacillus casei establishment in the gut. Proceedings of the National Academy of Sciences of the United States of America **111**(30), 3101–9 (2014). doi:10.1073/pnas.1411883111

[8] Lee, S.M., Donaldson, G.P., Mikulski, Z., Boyajian, S., Ley, K., Mazmanian, S.K.: Bacterial colonization factors control specificity and stability of the gut microbiota. Nature **501**(7467), 426–9 (2013). doi:10.1038/nature12447

[9] Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., González, A., Fontana, L., Henrissat, B., Knight, R., Gordon, J.I.: Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science (New York, N.Y.) **332**(6032), 970–4 (2011). doi:10.1126/science.1198719

[10] Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F.D., Lewis, J.D.: Linking long-term dietary patterns with gut microbial enterotypes. Science (New York, N.Y.) **334**(6052), 105–8 (2011). doi:10.1126/science.1208344

[11] Denou, E., Lolmède, K., Garidou, L., Pomie, C., Chabo, C., Lau, T.C., Fullerton, M.D., Nigro, G., Zakaroff-Girard, A., Luche, E., Garret, C., Serino, M., Amar, J., Courtney, M., Cavallari, J.F., Henriksbo, B.D., Barra, N.G., Foley, K.P., McPhee, J.B., Duggan, B.M., O'Neill, H.M., Lee, A.J., Sansonetti, P., Ashkar, A.A., Khan, W.I., Surette, M.G., Bouloumié, A., Steinberg, G.R., Burcelin, R., Schertzer, J.D.: Defective NOD2 peptidoglycan sensing promotes diet-induced inflammation, dysbiosis, and insulin resistance. EMBO molecular medicine **7**(3), 259–74 (2015). doi:10.15252/emmm.201404169

[12] Vijay-Kumar, M., Sanders, C.J., Taylor, R.T., Kumar, A., Aitken, J.D., Sitaraman, S.V., Neish, A.S., Uematsu, S., Akira, S., Williams, I.R., Gewirtz, A.T.: Deletion of TLR5 results in spontaneous colitis in mice. Journal of Clinical Investigation **117**(12), 3909–21 (2007). doi:10.1172/JCI33084

[13] Dethlefsen, L., Huse, S., Sogin, M.L., Relman, D.A.: The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS biology **6**(11), 280 (2008). doi:10.1371/journal.pbio.0060280

[14] Cox, L.M., Yamanishi, S., Sohn, J., Alekseyenko, A.V., Leung, J.M., Cho, I., Kim, S.G., Li, H., Gao, Z., Mahana, D., Zárate Rodriguez, J.G., Rogers, A.B., Robine,

N., Loke, P., Blaser, M.J.: Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. Cell **158**(4), 705–21 (2014). doi:10.1016/j.cell.2014.05.052

[15] Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J., Zhang, M., Oh, P.L., Nehrenberg, D., Hua, K., Kachman, S.D., Moriyama, E.N., Walter, J., Peterson, D.A., Pomp, D.: Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences of the United States of America **107**(44), 18933–8 (2010). doi:10.1073/pnas.1007028107

[16] Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature **486**(7402), 207–14 (2012). doi:10.1038/nature11234

[17] 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., Egholm, M., Henrissat, B., Heath, A.C., Knight, R., Gordon, J.I.: A core gut microbiome in obese and lean twins. Nature **457**(7228), 480–4 (2009). doi:10.1038/nature07540

[18] Young, V.B., Knox, K.A., Schauer, D.B.: Cytolethal distending toxin sequence and activity in the enterohepatic pathogen Helicobacter hepaticus. Infection and immunity **68**(1), 184–91 (2000)

[19] Mazmanian, S.K., Round, J.L., Kasper, D.L.: A microbial symbiosis factor prevents intestinal inflammatory disease. Nature **453**(7195), 620–5 (2008). doi:10.1038/nature07008

[20] Haiser, H.J., Gootenberg, D.B., Chatman, K., Sirasani, G., Balskus, E.P., Turnbaugh, P.J.: Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science (New York, N.Y.) **341**(6143), 295–8 (2013). doi:10.1126/science.1235872

[21] Wallace, B.D., Wang, H., Lane, K.T., Scott, J.E., Orans, J., Koo, J.S., Venkatesh, M., Jobin, C., Yeh, L.-A., Mani, S., Redinbo, M.R.: Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. Science (New York, N.Y.) **330**(6005), 831–5 (2010). doi:10.1126/science.1191175

[22] Soneson, C., Delorenzi, M.: A comparison of methods for differential expression analysis of RNA-seq data. BMC bioinformatics **14**, 91 (2013). doi:10.1186/1471-2105-14-91

[23] Yu, D., Huber, W., Vitek, O.: Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics (Oxford, England) **29**(10), 1275–82 (2013). doi:10.1093/bioinformatics/btt143

[24] Vallejos, C.A., Marioni, J.C., Richardson, S.: BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS computational biology **11**(6), 1004333 (2015). doi:10.1371/journal.pcbi.1004333

[25] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S.D., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J.: A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490**(7418), 55–60 (2012). doi:10.1038/nature11450

[26] Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., Bäckhed, F.: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature **498**(7452), 99–103 (2013). doi:10.1038/nature12198

[27] Nayfach, S., Bradley, P.H., Wyman, S.K., Laurent, T.J., Williams, A., Eisen, J.A., Pollard, K.S., Sharpton, T.J.: Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLoS computational biology **11**(11), 1004573 (2015). doi:10.1371/journal.pcbi.1004573

[28] Nayfach, S., Pollard, K.S.: Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome biology **16**, 51 (2015). doi:10.1186/s13059-015-0611-7

[29] Sauerwald, A., Zhu, W., Major, T.A., Roy, H., Palioura, S., Jahn, D., Whitman, W.B., Yates, J.R., Ibba, M., Söll, D.: RNA-dependent cysteine biosynthesis in archaea. Science (New York, N.Y.) **307**(5717), 1969–72 (2005). doi:10.1126/science.1108329

[30] Dridi, B., Henry, M., El Khéchine, A., Raoult, D., Drancourt, M.: High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. PloS one **4**(9), 7063 (2009). doi:10.1371/journal.pone.0007063

[31] Yanagisawa, T., Sumida, T., Ishii, R., Takemoto, C., Yokoyama, S.: A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P. Nature structural & molecular biology **17**(9), 1136–43 (2010). doi:10.1038/nsmb.1889

[32] Roy, H., Zou, S.B., Bullwinkle, T.J., Wolfe, B.S., Gilreath, M.S., Forsyth, C.J., Navarre, W.W., Ibba, M.: The tRNA synthetase paralog PoxA modifies elongation factor-P with (R)-β-lysine. Nature chemical biology **7**(10), 667–9 (2011). doi:10.1038/nchembio.632

[33] Peekhaus, N., Conway, T.: What's for dinner?: Entner-Doudoroff metabolism in Escherichia coli. Journal of bacteriology **180**(14), 3495–502 (1998)

[34] Spencer, M.E., Guest, J.R.: Isolation and properties of fumarate reductase mutants of Escherichia coli. Journal of bacteriology **114**(2), 563–70 (1973)

[35] Coburn, B., Sekirov, I., Finlay, B.B.: Type III secretion systems and disease. Clinical microbiology reviews **20**(4), 535–49 (2007). doi:10.1128/CMR.00013-07

[36] Coulthurst, S.J.: The Type VI secretion system - a widespread and versatile cell targeting system. Research in microbiology **164**(6), 640–54. doi:10.1016/j.resmic.2013.03.017

[37] Chatzidaki-Livanis, M., Geva-Zatorsky, N., Comstock, L.E.: Bacteroides fragilis type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species. Proceedings of the National Academy of Sciences of the United States of America **113**(13), 3627–32 (2016). doi:10.1073/pnas.1522510113

[38] Wexler, A.G., Bao, Y., Whitney, J.C., Bobay, L.-M., Xavier, J.B., Schofield, W.B., Barry, N.A., Russell, A.B., Tran, B.Q., Goo, Y.A., Goodlett, D.R., Ochman, H., Mougous, J.D., Goodman, A.L.: Human symbionts inject and neutralize antibacterial toxins to persist in the gut. Proceedings of the National Academy of Sciences **113**(13), 201525637 (2016). doi:10.1073/pnas.1525637113

[39] Cao, T.B., Saier, M.H.: The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. Biochimica et biophysica acta **1609**(1), 115–25 (2003)

[40] Wu, D.: personal communication

[41] Hooper, L.V., Midtvedt, T., Gordon, J.I.: How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annual review of nutrition **22**, 283–307 (2002). doi:10.1146/annurev.nutr.22.011602.092259

[42] Benjdia, A., Martens, E.C., Gordon, J.I., Berteau, O.: Sulfatases and a radical S-adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the prominent human gut symbiont, Bacteroides thetaiotaomicron. The Journal of biological chemistry **286**(29), 25973–82 (2011). doi:10.1074/jbc.M111.228841

[43] Ulmer, J.E., Vilén, E.M., Namburi, R.B., Benjdia, A., Beneteau, J., Malleron, A., Bonnaffé, D., Driguez, P.-A., Descroix, K., Lassalle, G., Le Narvor, C., Sandström, C., Spillmann, D., Berteau, O.: Characterization of glycosaminoglycan (GAG) sulfatases from the human gut symbiont Bacteroides thetaiotaomicron reveals the first GAG-specific bacterial endosulfatase. The Journal of biological chemistry **289**(35), 24289–303 (2014). doi:10.1074/jbc.M114.573303

[44] Raghavan, V., Groisman, E.A.: Species-specific dynamic responses of gut bacteria to a mammalian glycan. Journal of bacteriology **197**(9), 1538–48 (2015). doi:10.1128/JB.00010-15

[45] Greenblum, S., Carr, R., Borenstein, E.: Extensive strain-level copy-number variation across human gut microbiome species. Cell **160**(4), 583–94 (2015). doi:10.1016/j.cell.2014.12.038

[46] Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N.: MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods **12**(10), 902–903 (2015). doi:10.1038/nmeth.3589

[47] Shin, N.-R., Whon, T.W., Bae, J.-W.: Proteobacteria: microbial signature of dysbiosis in gut microbiota. Trends in biotechnology **33**(9), 496–503 (2015). doi:10.1016/j.tibtech.2015.06.011

[48] Mukhopadhya, I., Hansen, R., El-Omar, E.M., Hold, G.L.: IBD-what role do Proteobacteria play? Nature reviews. Gastroenterology & hepatology **9**(4), 219–30 (2012). doi:10.1038/nrgastro.2012.14

[49] Garrett, W.S., Gallini, C.A., Yatsunenko, T., Michaud, M., DuBois, A., Delaney, M.L., Punit, S., Karlsson, M., Bry, L., Glickman, J.N., Gordon, J.I., Onderdonk, A.B., Glimcher, L.H.: Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell host & microbe **8**(3), 292–300 (2010). doi:10.1016/j.chom.2010.08.004

[50] Carvalho, F.A., Koren, O., Goodrich, J.K., Johansson, M.E.V., Nalbantoglu, I., Aitken, J.D., Su, Y., Chassaing, B., Walters, W.A., González, A., Clemente, J.C., Cullen-

der, T.C., Barnich, N., Darfeuille-Michaud, A., Vijay-Kumar, M., Knight, R., Ley, R.E., Gewirtz, A.T.: Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice. Cell host & microbe **12**(2), 139–52 (2012). doi:10.1016/j.chom.2012.07.004

[51] Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., Gordon, J.I.: Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America **102**(31), 11070–5 (2005). doi:10.1073/pnas.0504978102

[52] Nayfach, S., Pollard, K.S.: Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. Technical report (nov 2015). doi:10.1101/031757. http://biorxiv.org/content/early/2015/11/14/031757.abstract

[53] Ma, L., Terwilliger, A., Maresso, A.W.: Iron and zinc exploitation during bacterial pathogenesis. Metallomics : integrated biometal science **7**(12), 1541–54 (2015). doi:10.1039/c5mt00170f

[54] Wallden, K., Rivera-Calzada, A., Waksman, G.: Type IV secretion systems: versatility and diversity in function. Cellular microbiology **12**(9), 1203–12 (2010). doi:10.1111/j.1462-5822.2010.01499.x

[55] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C.: Metagenomic biomarker discovery and explanation. Genome biology **12**(6), 60 (2011). doi:10.1186/gb-2011-12-6-r60

[56] Brennecke, P., Anders, S., Kim, J.K., Ko\l{}odziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G.: Accounting for technical noise in single-cell RNA-seq experiments. Nature methods **10**(11), 1093–5 (2013). doi:10.1038/nmeth.2645

[57] Dumitrascu, B., Darnell, G., Ayroles, J., Engelhardt, B.E.: A Bayesian test to identify variance effects. 1512.01616. http://arxiv.org/abs/1512.01616

[58] Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., Knight, R.: Diversity, stability and resilience of the human gut microbiota. Nature **489**(7415), 220–30 (2012). doi:10.1038/nature11550

[59] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. Nucleic acids research **32**(Database issue), 277–80 (2004). doi:10.1093/nar/gkh063

[60] Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America **100**(16), 9440–5 (2003). doi:10.1073/pnas.1530509100

[61] Storey, J.D., Bass, A.J., Dabney, A., Robinson, D.: qvalue: Q-value estimation for false discovery rate control (2015). http://github.com/jdstorey/qvalue

[62] Pollard, K., van der Laan, M.: Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data (2003). http://biostats.bepress.com/ucbbiostat/paper121

[63] Pollard, K.S., van der Laan, M.J.: Choice of a null distribution in resampling-based multiple testing. Journal of Statistical Planning and Inference **125**(1-2), 85–100 (2004). doi:10.1016/j.jspi.2003.07.019

[64] Lloyd-Smith, J.O.: Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PloS one **2**(2), 180 (2007). doi:10.1371/journal.pone.0000180

[65] Sievers, F., Higgins, D.G.: Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In: Methods in Molecular Biology (Clifton, N.J.) vol. 1079, pp. 105–116 (2014). http://www.ncbi.nlm.nih.gov/pubmed/24170397 http://link.springer.com/10.1007/978-1-62703-646-7_6

[66] Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one **5**(3), 9490 (2010). doi:10.1371/journal.pone.0009490

[67] Kim, S.: ppcor: Partial and Semi-Partial (Part) Correlation (2015). http://cran.r-project.org/package=ppcor

[68] Zhao, Y., Tang, H., Ye, Y.: RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics (Oxford, England) **28**(1), 125–6 (2012). doi:10.1093/bioinformatics/btr595

[69] Wu, D., Jospin, G., Eisen, J.A.: 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 **8**(10), 77033 (2013). doi:10.1371/journal.pone.0077033

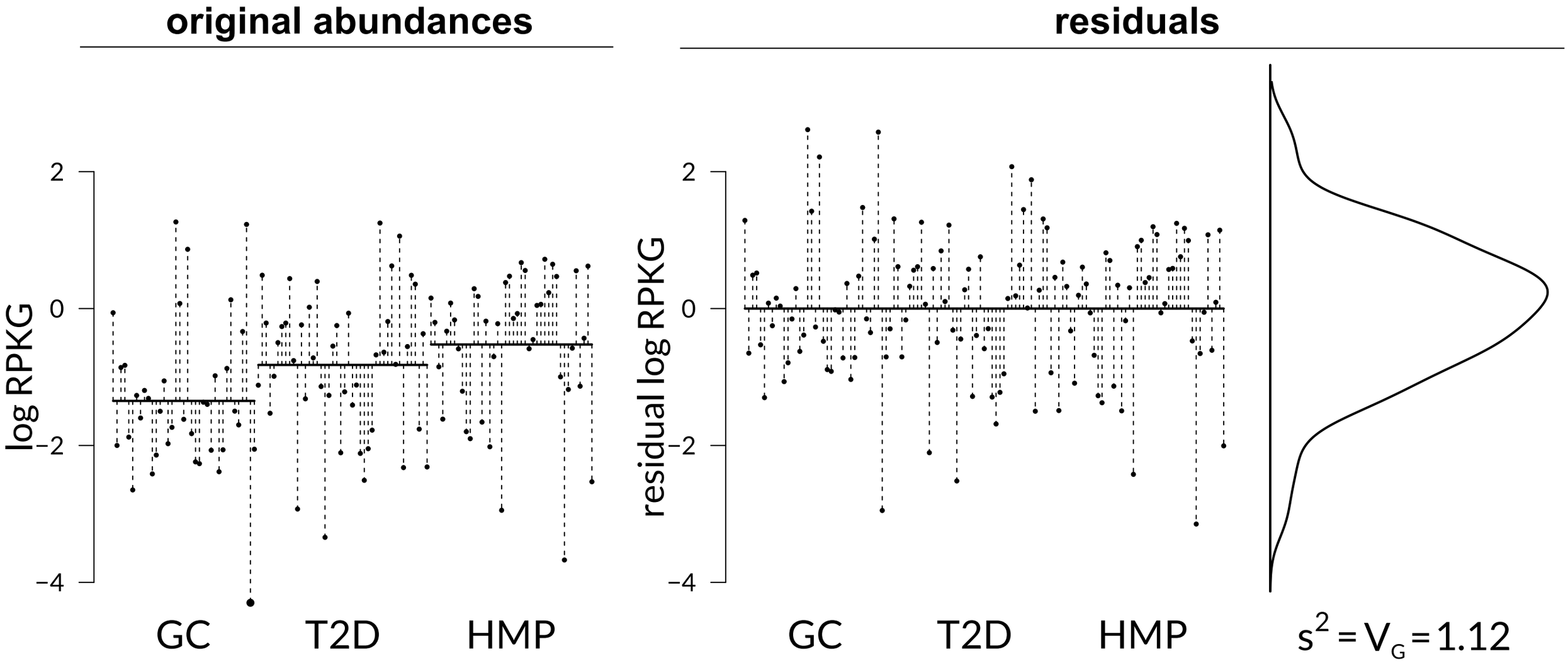
[70] Manor, O., Borenstein, E.: MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biology **16**(1), 53 (2015). doi:10.1186/s13059-015-0610-8

[71] Wagner, C., de Saizieu Ad, A., Schönfeld, H.-J., Kamber, M., Lange, R., Thompson, C.J., Page, M.G.: Genetic analysis and functional characterization of the Streptococcus pneumoniae vic operon. Infection and immunity **70**(11), 6121–8 (2002)

A

waaL family

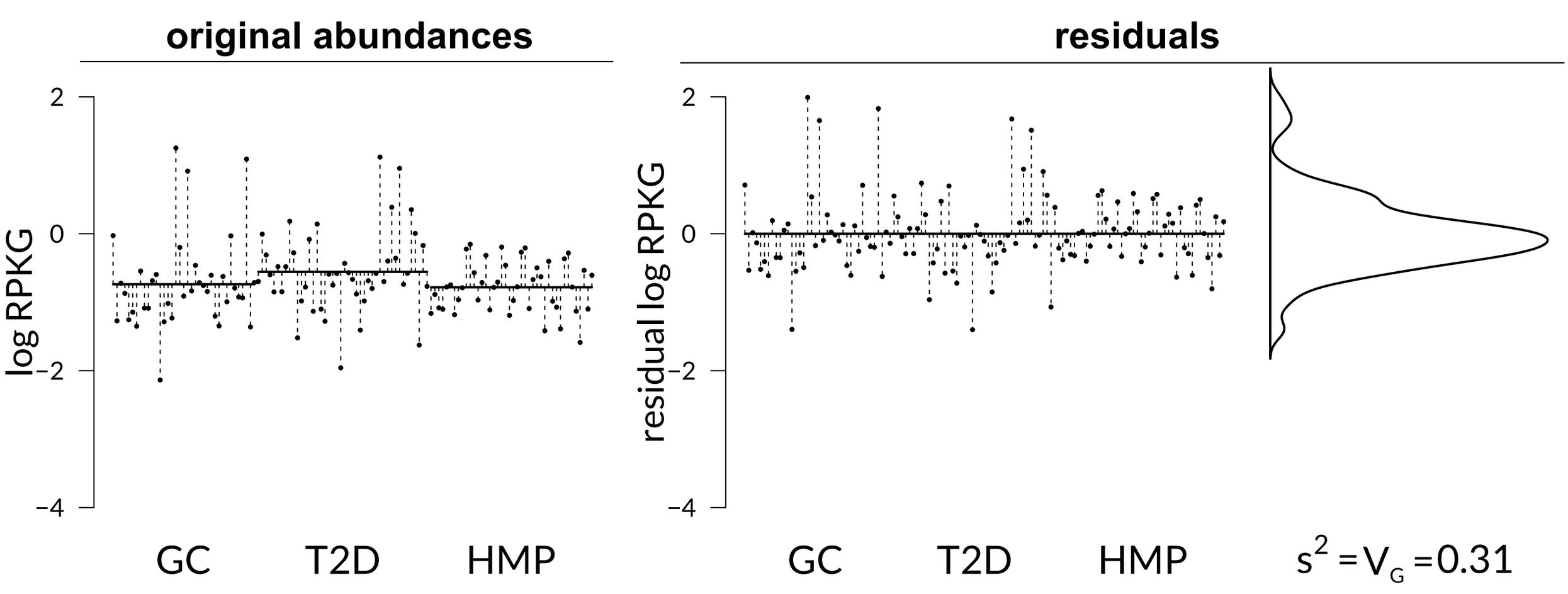
lipopolysaccharide O-antigen ligase

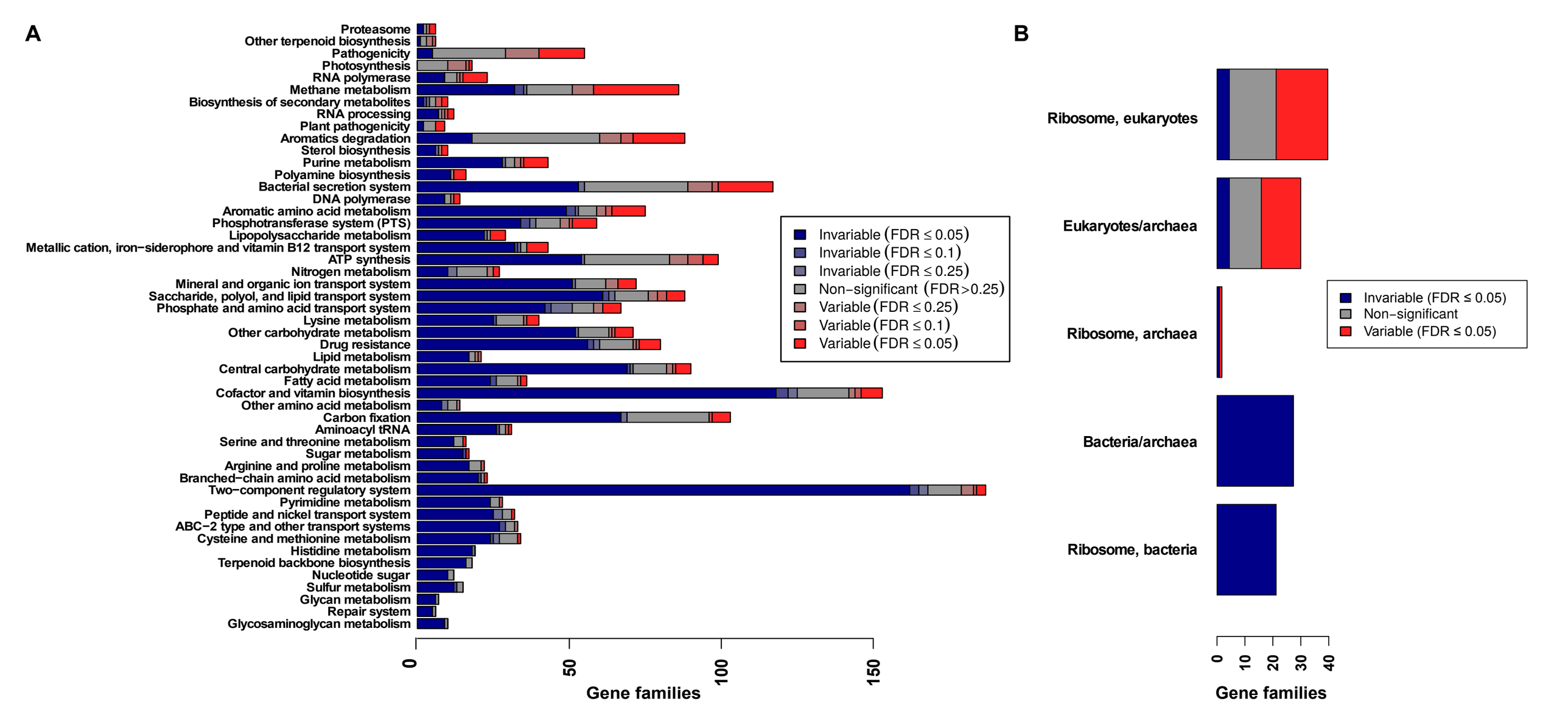


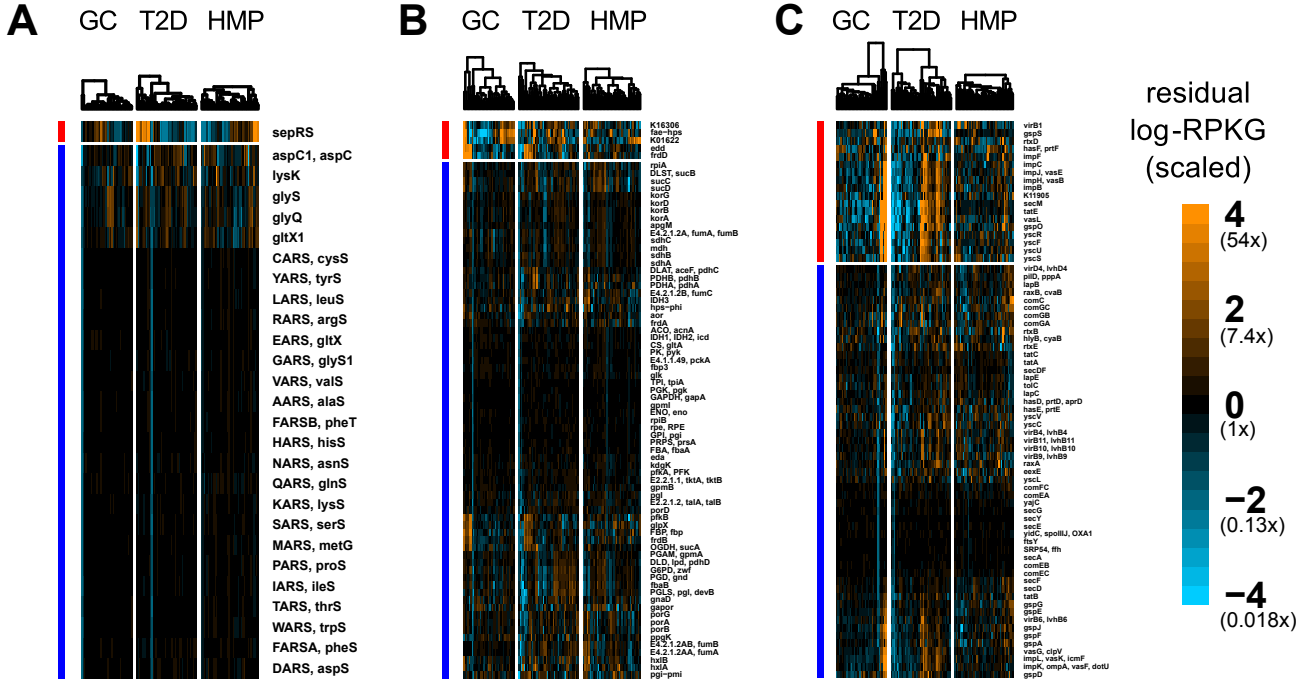
B

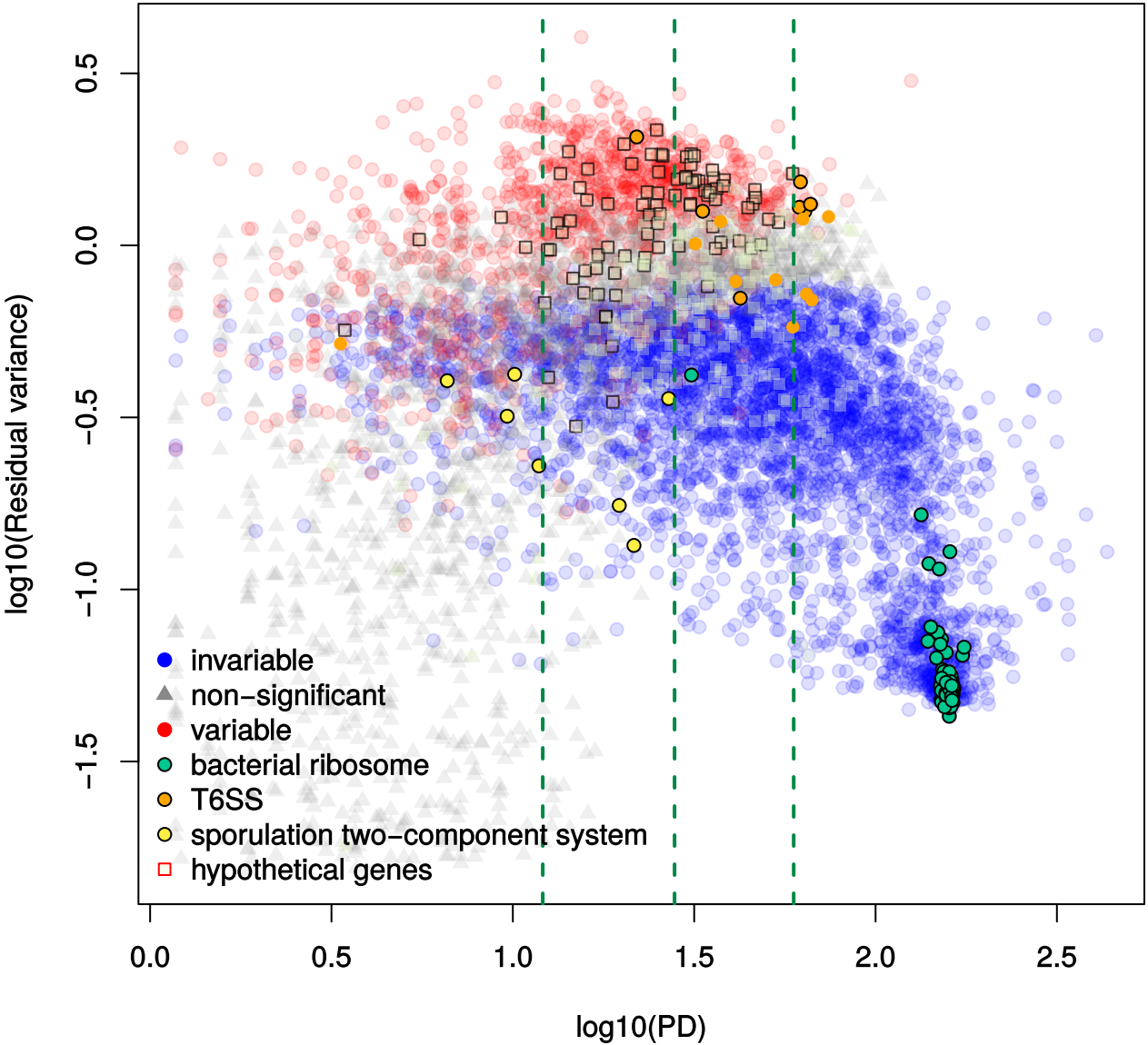
vicR family

two-component OmpR family transcriptional regulator



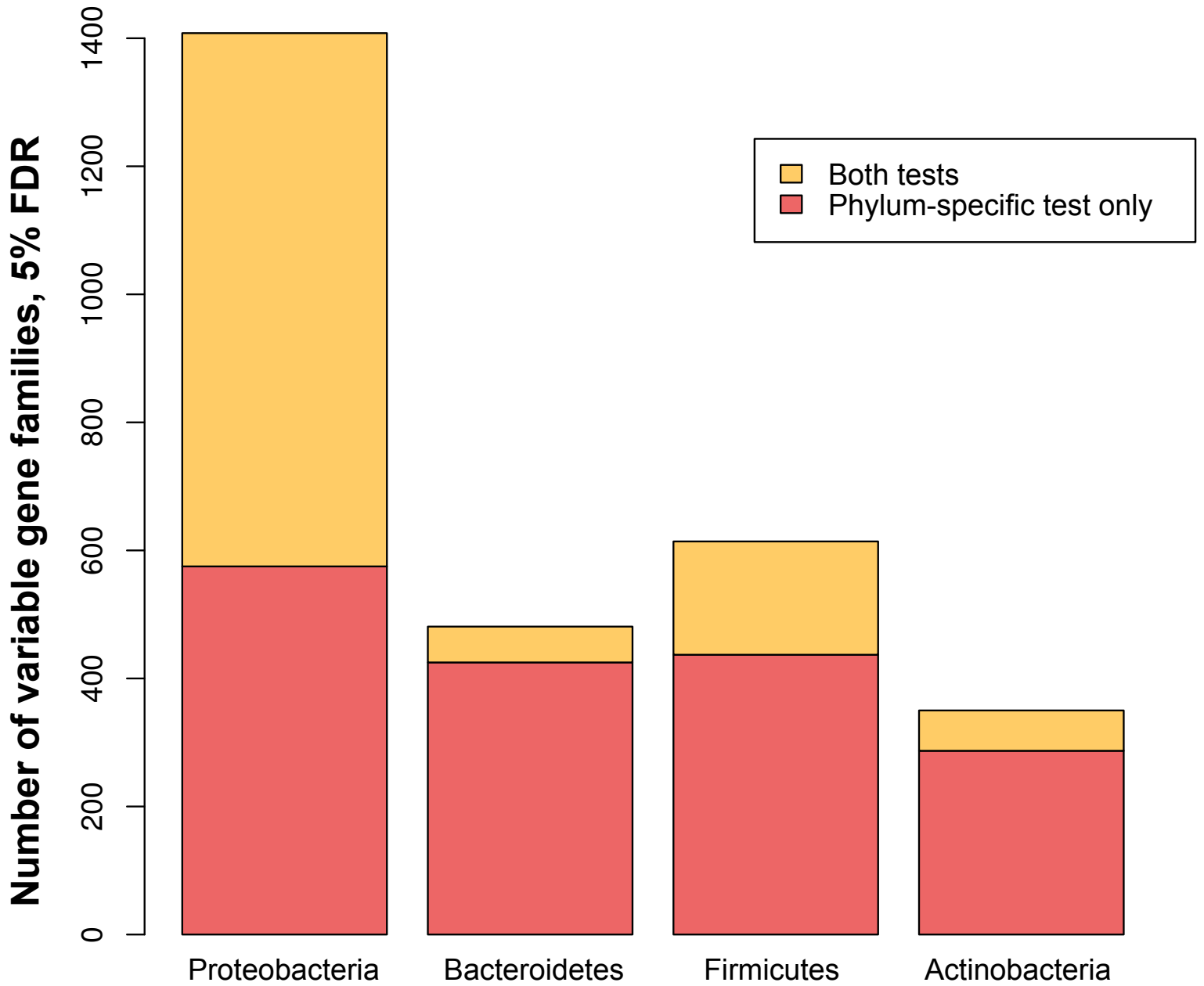

A
Proteasome
Other terpenoid biosynthesis
Pathogenicity
Photosynthesis
RNA polymerase
Methane metabolism
Biosynthesis of secondary metabolites
RNA processing
Plant pathogenicity
Aromatics degradation
Sterol biosynthesis
Purine metabolism
Polyamine biosynthesis
Bacterial secretion system
DNA polymerase
Aromatic amino acid metabolism
Phosphotransferase system (PTS)
Lipopolysaccharide metabolism
Metallic cation, iron−siderophore and vitamin B12 transport system
ATP synthesis
Nitrogen metabolism
Mineral and organic ion transport system
Saccharide, polyol, and lipid transport system
Phosphate and amino acid transport system
Lysine metabolism
Other carbohydrate metabolism
Drug resistance
Lipid metabolism
Central carbohydrate metabolism
Fatty acid metabolism
Cofactor and vitamin biosynthesis
Other amino acid metabolism
Carbon fixation
Aminoacyl tRNA
Serine and threonine metabolism
Sugar metabolism
Arginine and proline metabolism
Branched−chain amino acid metabolism
Two−component regulatory system
Pyrimidine metabolism
Peptide and nickel transport system
ABC−2 type and other transport systems
Cysteine and methionine metabolism
Histidine metabolism
Terpenoid backbone biosynthesis
Nucleotide sugar
Sulfur metabolism
Glycan metabolism
Repair system
Glycosaminoglycan metabolism
Invariable (FDR ≤ 0.05)
Invariable (FDR ≤ 0.1)
Invariable (FDR ≤ 0.25)
Non−significant (FDR>0.25)
Variable (FDR ≤ 0.25)
Variable (FDR ≤ 0.1)
Variable (FDR ≤ 0.05)
0
50
100
150
Gene families
B
Ribosome, eukaryotes
Eukaryotes/archaea
Ribosome, archaea
Bacteria/archaea
Ribosome, bacteria
Invariable (FDR ≤ 0.05)
Non−significant
Variable (FDR ≤ 0.05)
0
10
20
30
40
Gene families


A
GC
T2D
HMP
sepRS
aspC1, aspC
lysK
glyS
glyQ
gltX1
CARS, cysS
YARS, tyrS
LARS, leuS
RARS, argS
EARS, gltX
GARS, glyS1
VARS, valS
AARS, alaS
FARSB, pheT
HARS, hisS
NARS, asnS
QARS, glnS
KARS, lysS
SARS, serS
MARS, metG
PARS, proS
IARS, ileS
TARS, thrS
WARS, trpS
FARSA, pheS
DARS, aspS
B
GC
T2D
HMP
K16306
fae−hps
K01622
edd
frdD
rpiA
DLST, sucB
sucC
sucD
korG
korD
korB
korA
apgM
E4.2.1.2A, fumA, fumB
sdhC
mdh
sdhB
sdhA
DLAT, aceF, pdhC
PDHB, pdhB
PDHA, pdhA
E4.2.1.2B, fumC
IDH3
hps−phi
aor
frdA
ACO, acnA
IDH1, IDH2, icd
CS, gltA
PK, pyk
E4.1.1.49, pckA
fbp3
glk
TPI, tpiA
PGK, pgk
GAPDH, gapA
gpmI
ENO, eno
rpiB
rpe, RPE
GPI, pgi
PKPS, prsA
FBA, fbaA
eda
kdgK
pfkA, PFK
E2.2.1.1, tktA, tktB
gpmB
pgl
E2.2.1.2, talA, talB
porD
pfkB
glpX
FBP, fbp
frdB
OGDH, sucA
PGAM, gpmA
DLD, lpd, pdhD
G6PD, zwf
PGD, gnd
fbaB
PGLS, pgl, devB
gnaD
gapor
porG
porA
porB
ppgK
E4.2.1.2AB, fumB
E4.2.1.2AA, fumA
hxlB
hxlA
pgi−pmi
C
GC
T2D
HMP
virB1
gspS
rtxD
hasF, prtF
impF
impC
impJ, vasE
impH, vasB
impB
K11905
secM
tatE
vasL
gspD
yscR
yscF
yscU
yscS
virD4, lvhD4
pilD, pppA
lapB
ravB, cvaB
comC
comGC
comGB
comGA
rtxB
hlyB, cyaB
rtxE
tatC
tatA
secDF
lspE
tolC
lapC
hasD, prtD, aprD
hasE, prtE
yscV
yscC
virB4, lvhB4
virB11, lvhB11
virB10, lvhB10
virB9, lvhB9
raxA
exeE
yscL
comFC
comEA
yqjC
secG
secY
secE
yidC, spoIIIJ, OXA1
ftsY
SRP54, ffh
secA
comEB
comEC
secF
secD
tatB
gspG
gspE
virB6, lvhB6
gspJ
gspF
gspA
vasG, clpV
impL, vasK, icmF
impK, ompA, vasF, dotU
gspD
residual
log-RPKG
(scaled)
4
(54x)
2
(7.4x)
0
(1x)
−2
(0.13x)
−4
(0.018x)


log10(Residual variance)
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
log10(PD)
invariable
non−significant
variable
bacterial ribosome
T6SS
sporulation two−component system
hypothetical genes

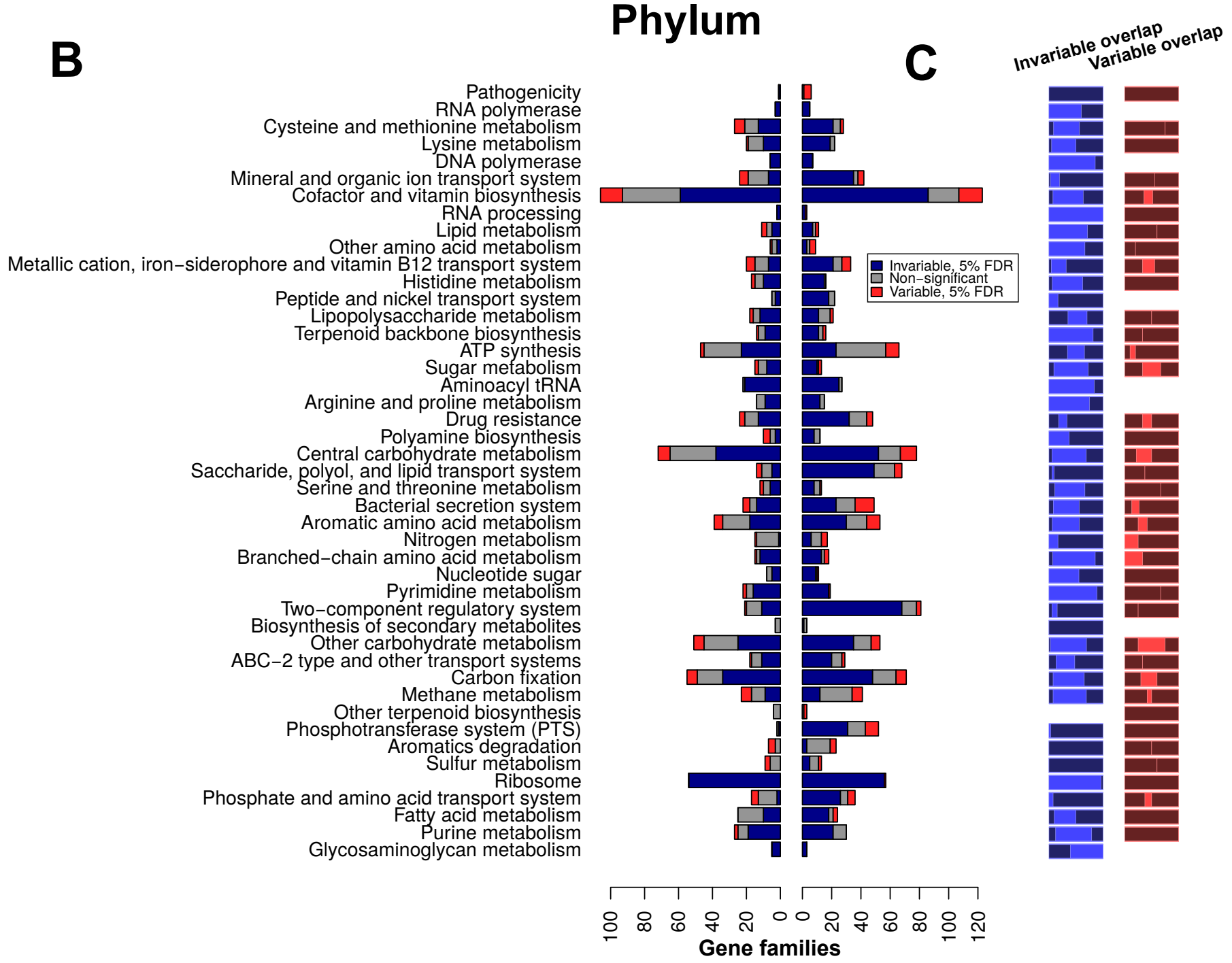
Fraction of hits to any phylum
0.0
0.1
0.2
0.3
0.4
0.5
Invariable
Non-significant
Variable
*** Firmicutes
*** Bacteroidetes
*** Proteobacteria
*** Actinobacteria
** Euryarchaeota
Viruses (unnamed)
*** Ascomycota
*** Verrucomicrobia
*** Candidatus Saccharibacteria
Fusobacteria
Synergistetes
Tenericutes
Spirochaetes
Deinococcus–Thermus

A


Number of variable gene families, 5% FDR
0
200
400
600
800
1000
1200
1400
Both tests
Phylum-specific test only
Proteobacteria
Bacteroidetes
Firmicutes
Actinobacteria

B

C


Phylum
Pathogenicity
RNA polymerase
Cysteine and methionine metabolism
Lysine metabolism
DNA polymerase
Mineral and organic ion transport system
Cofactor and vitamin biosynthesis
RNA processing
Lipid metabolism
Other amino acid metabolism
Metallic cation, iron−siderophore and vitamin B12 transport system
Histidine metabolism
Peptide and nickel transport system
Lipopolysaccharide metabolism
Terpenoid backbone biosynthesis
ATP synthesis
Sugar metabolism
Aminoacyl tRNA
Arginine and proline metabolism
Drug resistance
Polyamine biosynthesis
Central carbohydrate metabolism
Saccharide, polyol, and lipid transport system
Serine and threonine metabolism
Bacterial secretion system
Aromatic amino acid metabolism
Nitrogen metabolism
Branched−chain amino acid metabolism
Nucleotide sugar
Pyrimidine metabolism
Two−component regulatory system
Biosynthesis of secondary metabolites
Other carbohydrate metabolism
ABC−2 type and other transport systems
Carbon fixation
Methane metabolism
Other terpenoid biosynthesis
Phosphotransferase system (PTS)
Aromatics degradation
Sulfur metabolism
Ribosome
Phosphate and amino acid transport system
Fatty acid metabolism
Purine metabolism
Glycosaminoglycan metabolism
Invariable, 5% FDR
Non−significant
Variable, 5% FDR
100
80
60
40
20
0
0
20
40
60
80
100
120
Gene families
Invariable overlap
Variable overlap