Phylogeny-corrected identification of microbial gene families relevant to human gut colonization

Patrick H. Bradley, Stephen Nayfach, Katherine S. Pollard

September 15, 2017

Abstract

The mechanisms by which different microbes colonize the healthy human gut versus free-living communities, other body sites, or the gut in disease states remain largely unknown. Identifying microbial genes influencing fitness in the gut could lead to new ways to engineer probiotics or disrupt pathogenesis. We propose a statistical approach to this problem that measures the association between having a gene and the probability that a species is present in the gut microbiome. The challenge is that closely related species tend to be jointly present or absent in the microbiome and also share many genes, only a subset of which are involved in gut adaptation. We show that this phylogenetic correlation indeed leads to many false discoveries and propose phylogenetic linear regression as a powerful solution. To apply this method across the bacterial tree of life, where most species have not been experimentally phenotyped, we used metagenomes from hundreds of people to quantify each species' prevalence in and specificity for the gut microbiome. This analysis revealed thousands of genes potentially involved across species in adaptation to the gut, including many novel candidates as well as processes known to contribute to fitness of gut bacteria, such as acid tolerance in Bacteroidetes and sporulation in Firmicutes. We also found microbial genes associated with a preference for the gut over other body sites, which were significantly enriched for genes linked to fitness in an in vivo competition experiment. Finally, we identified gene families associated with higher prevalence in patients with Crohn's disease, including Proteobacterial genes involved in conjugation and fimbria regulation, processes previously linked to inflammation. These gene targets may represent new avenues for modulating host colonization and disease. Our strategy of combining metagenomics with phylogenetic modeling is general and can be used to identify genes associated with adaptation to any environment.

Author Summary

Why do certain microbes and not others colonize our gut, and why do they differ between healthy and sick people? One explanation is the genes in their genomes. If we can find microbial genes involved in gut adaptation, we may be able to keep out pathogens and encourage the growth of beneficial microbes. One could look for genes that were present more often in prevalent microbes, and less often in rare ones. However, this ignores that similar species may have related phenotypes simply because of common ancestry. To solve this problem, we used a method from ecology that accounts for phylogenetic relatedness. We first calculated gut prevalence phenotypes for thousands of species using a compedium of shotgun sequencing data, then tested for gene associations. We found genes that are associated with overall gut prevalence, with a preference for the gut over other body sites, and with the gut in Crohn's disease vs. health. Many of these findings have biological plausibility based on existing literature. We also showed agreement with the results of a previously published high-throughput screen of bacterial gene knockouts in mice. These results, and this type of analysis, may eventually lead to new strategies for maintaining gut health.

Short title

Phylogenetic modeling of gut colonization

1 Background

Microbes that colonize the human gastrointestinal (GI) tract have a wide variety of effects on their hosts, ranging from beneficial to harmful. Increasing evidence shows that commensal gut microbes are responsible for training and modulating the immune system [1, 2], protecting against inflammation [3] and pathogen invasion (reviewed in Sassone-Corsi and Raffatellu [4]), affecting GI motility [5], maintaining the intestinal barrier [6], and potentially even affecting mood [7]. In contrast, pathogens (and conditionally-pathogenic microbes, or "pathobionts") can induce and worsen inflammation [8, 9], increase the risk of cancer in mouse models [10], and cause potentially life-threatening infections [11]. Additionally, the transplantation of microbes from a healthy host (fecal microbiota transplant, or FMT) is also a highly effective therapy for some gut infections [12], although it is still an active area of investigation why certain microbes from the donor persist long-term and others do not [13], and how pre-existing inflammatory disease affects FMT efficacy [14]. Which microbes are able to persist in the GI tract, and why some persist instead of others, is therefore a question with consequences that directly impact human health.

Because of this, we are interested in the specific mechanisms by which microbes colonize the gut, avoiding other potential fates such as being killed in the harsh stomach environment, simply passing through the GI tract transiently, or being outcompeted by other gut microbes. Understanding these mechanisms could yield opportunities to design better probiotics and to prevent invasion of the gut community by pathogens. In particular, creating new therapies, whether those are drugs, engineered bacterial strains, or rationally designed communities, will likely require an understanding of gut colonization at the level of individual microbial genes. We also anticipate that these mechanisms may vary in health vs. disease, since, for example, different selective pressures are known to be present in inflamed versus healthy guts [15, 16].

One approach that has been used to link genetic features to a phenotype is to correlate the two using observational data. Most typically, this approach is applied in the form of genome-wide association mapping, in which phenotypes are correlated with genetic markers across individuals in a population. While we are interested in comparing phenotypes and genetic features *across,* rather than within species, the approach we take in this paper is conceptually similar. In order to perform association mapping, it is necessary to account for population structure, that is, dependencies resulting from common ancestry; otherwise, spurious discoveries can be made in genome-wide association studies [17]. Analogously, we expected it to be important to choose a method that can account for the confounding effect of phylogeny when testing for associations across species.

There is increasing interest in using phylogenetic information to make better inferences about associations between microbes and quantities of interest. For example, co-conservation patterns of genes ("correlogs") have been used to assign functions to microbial genes [18], and genome-wide association studies have been applied within a genus of soil bacteria [19]. Recent publications have also described techniques that use information from the taxonomic tree to more accurately link clades in compositional taxonomic data to covariates [20, 21, 22]. However, so far, only one study has attempted to associate genes with a preference for the gut [23]. That study introduced a valuable method based on UniFrac and gene-count distances, which compares how well gut- vs. non-gut-associated microbes cluster on the species tree compared to a composite gene tree. This study also provides an important insight in the form of evidence of convergence of glycoside hydrolase and glycosyltransferase repertoires among gut bacteria, suggesting horizontal gene transfer within the gut community to deal with a common evolutionary pressure. The method described in that study, though, requires a binary phenotype of gut presence vs. absence. Deciding which microbes are "gut" vs. "non-gut" requires manual curation and can be somewhat subjective, as microbes have a continuous range of prevalences and can appear in multiple environments; this binarization could also potentially decrease power by excluding microbes with intermediate phenotypes. The method also requires multiple sequence alignments and trees to be built for every gene family under analysis, which are computationally intensive to generate over a large set of genomes.

We take a complementary approach and use a flexible technique, known as phylogenetic linear modeling, to detect associations between microbial genotype and phenotype while accounting for the fact that microbes are related to one another by vertical descent. Phylogenetic linear models have an extensive history in the ecology literature dating back to seminal works by Felsenstein [24] and Grafen [25]. However, despite their power, genome-scale applications of these models are still few in number [26] and have typically been used to relate traits of macroorganisms (e.g., anole lizards [27]) to their genotypes. While there is a growing

appreciation for the need to explicitly account for phylogeny in microbial community analyses [26, 28], we believe ours is the first study to apply this class of methods to microbiome data.

Our approach to accounting for phylogenetic relationships is general and could be applied to measure association of any quantitative trait with genotypes or other binary or quantitative characteristics. In this study, we focus on phenotypes related to the ability of bacteria to colonize the human gut: 1. overall prevalence in the guts of hosts from a specific population (e.g., post-industrialized countries), which we expect to capture ease of transmission, how cosmopolitan microbes are, and how efficiently they colonize the gut; 2. a preference for the gut over other human body sites in the same hosts, which we expect to capture gut colonization more specifically; and 3. a preference for the gut in disease (e.g., Crohn's disease) versus health. We present a novel method to estimate these quantitative phenotypes for thousands of bacterial species directly from existing shotgun metagenomics data, both obviating the need for us to draw a cutoff between "gut" and "non-gut" microbes, and also giving us the necessary power to detect associations. Coupling these phenotype estimates with phylogenetic linear models, we generate a compendium of thousands of bacterial genes whose functions may be involved in colonizing the human gut.

2 Results

We present a phylogeny-aware method for modeling associations between the presence of specific genes in bacterial genomes and quantitative traits that measure how common these species are in the human microbiome. To apply phylogenetic linear modeling to the microbiome, we needed to solve three problems. First, we had to show that these models controlled false positives and had reasonable power on large bacterial phylogenies. Second, we needed to develop estimators that captured meaningful traits related to bacterial colonization of humans for thousands of diverse bacterial species, most of which had not been directly phenotyped. The third problem was to estimate genotypes (e.g., gene presence/absence) for each species. The analysis framework we describe is quite general and could be easily extended to link other traits to genotypes across the tree of life.

2.1 Phylogenetic linear models solve the problem of high false positive rates when testing for associations on bacterial phylogenies

To test for associations between continuous traits and binary phenotypes across species, we use models with the following form:

$$Y_T = \beta_0 + \beta_{1,g} g_T + \epsilon_T \quad (1)$$

Y_T is the quantitative trait value of species T, β_0 is a baseline level of Y_T, $\beta_{1,g}$ is the effect of gene g on Y_T, g_T is an indicator variable (0 if the gene is absent in species T and 1 if present), and ϵ_T is the remaining unmodeled variation in Y_T. The distribution of the residuals ϵ_T is the key difference between standard and phylogenetic linear models. In the standard model, the residuals are assumed to be independent and normally distributed. In the phylogenetic model, however, the residuals covary, with more closely-related species having greater covariance (see Methods).

To explore the potential pitfalls of failing to correct for phylogenetic structure in cross-species association tests, we generated a species tree for thousands of bacteria with genome sequences (see Methods). In order to have a consistent operational definition of a microbial species, we used a set of previously defined bacterial taxonomic units with approximately 95% pairwise average nucleotide identity across the entire genome [29]. The methods we describe can be applied to other taxonomic levels or with other species definitions. Using this species tree, we performed simulations for each of the four major bacterial phyla in the human gut (Bacteroidetes, Firmicutes, Proteobacteria, and Actinobacteria). Specifically, we generated continuous traits with phylogenetic signal along the species tree, and then, for each continuous trait, simulated a binary trait for each species that was correlated with the continuous trait to varying degrees including no association. We used levels of correlation spanning those we observe between prevalence of species in gut metagenomes and presence/absence of genes (see below).

We then fit phylogenetic and standard linear models to the simulated data and tested for a relationship between each binary trait and its corresponding continuous trait. For both standard and phylogenetic linear

models, separate models were fit for each of the four phyla. The results were used to estimate false positive rate (Type I error) and power (1 - Type II error) for the two methods across different effect sizes.

These analyses showed that standard linear models result in many false positive associations. When the binary trait was specified to be wholly uncorrelated (i.e., under the null), p-values from the linear model showed a strong anticonservative bias (Figure 1A-B, right) with many more significant p-values than expected under no correlation. In contrast, the phylogenetic linear model p-value distribution was flat and Type I error was controlled appropriately (Figure 1A-B, left). This means that at the same p-value threshold, linear models will identify many spurious relationships compared to phylogenetic linear models. Further, our simulations with non-zero associations showed that the phylogenetic model has high power when applied to gut bacterial phyla, even for small effect sizes (Figure 1C; see Methods). These results emphasize the importance of using models that account for phylogenetic relationships in cross-species association testing and demonstrate the feasibility of applying phylogenetic linear models to the human microbiome.

2.2 Estimating quantitative phenotypes from shotgun data

To apply phylogenetic linear modeling to the microbiome we sought to define meaningful traits for thousands of bacterial species, all of which have genome sequences but most of which have never been phenotyped. The prevalence and specificity of bacterial species in an environment, such as the human gut, are quantitative traits that we hypothesized could be estimated directly from shotgun metagenomics data. The precise taxonomic composition of a healthy gut microbiome can vary significantly from person to person, indicating that the ability of a microbe to colonize the gut is quantitative (and likely context-dependent, and stochastic). This trait can be conceptualized differently depending on which aspects of colonization one wishes to capture. We present metagenome-based estimators for two different types of colonization trait parameters. These are described in the context of our goal of studying the gut microbiome, but the approach is general and could be used to quantify how well a given genotype discriminates species found in or specific to any environment.

The first trait is the probability of observing a species in an environment, that is, its *overall prevalence.* Both genes relating to survival in the GI tract and genes relating to survival, persistence, and dispersal in the outside environment are expected to correlate with overall prevalence. Prevalence can be estimated by the frequency with which the species is observed in a sample from the environment, for example, using a logit transform to enable linear modeling and pseudocounts to avoid estimates of 0 or 1 (see Methods).

The second type of quantitative trait is the *environmental specificity* of a microbial species, which we define as the conditional probability that a sample is derived from a specific environment versus others, given that the species is present in the sample. This parameter captures the power of a given microbe as a marker to discriminate between two or more different environments, such as different body sites or types of hosts. This is distinct from its overall prevalence in the environment.

We developed estimators for two examples of environmental specificity for gut microbes. First, we considered a phenotype defined as the conditional probability that a given body site is the gut and not another body site, given that a particular species is present. The physical distance between body sites is much smaller than the distance between hosts, and microbes from one body site are likely to be transiently introduced to others. Hence, enrichment of a species in one body site over others is stronger evidence for selection (versus dispersal) than is overall prevalence in that body site alone. We estimate this parameter with a *body-site specificity score* that uses metagenomics data to measure how predictive a particular microbe is for the gut versus other body sites (e.g., skin, urogenital tract, oropharynx, or lung). The score is based on a maximum *a posteriori* (MAP) estimate of the conditional probability of a sample being from the gut given that a microbe is observed in the sample, and it utilizes Laplace regularization, as in the Bayesian lasso [30], to perform a type of L_1-norm penalization of parameter estimates (see Methods).

The second type of environmental specificity we considered is the conditional probability that a host has a disease given that a particular species is present. This *disease-specific specificity score* is estimated in a similar way to the body-site specificity score (see Methods). We focus on Crohn's disease, a type of inflammatory bowel disease known to be associated with dramatic shifts in the gut microbiota and in gut-immune interactions [31]. Genes associated with this disease-specific prevalence could illuminate differences in selective pressures between healthy vs. diseased gut environments.

2.3 Genes associated with species prevalence in healthy human gut metagenomes

We assembled a compendium of published DNA sequencing data from healthy human stool microbiomes across five studies in North America, Europe, and China (433 subjects total). Multiple replicates from the same individual were merged. Using the MIDAS 1.0 database and pipeline [29], we mapped metagenomic sequencing reads from each subject to a panel of phylogenetic marker genes, and from these, estimated species relative abundances. We then estimated the prevalence (probability of non-zero abundance) of each species across these subjects, weighting each study equally (see Methods). Finally, we determined gene presence for each species using its pangenome. This approach to genotyping could be elaborated to account for metagenomic sequencing data and assemblies in future work (see Discussion). Our analysis framework can also be applied to genotypes other than gene presence/absence (e.g., nucleotide or amino acid changes).

As expected, the most prevalent species overall included *Bacteroides vulgatus*, *Bacteroides ovatus*, and *Faecalibacterium prausnitzii*, while the least prevalent included halophiles and thermophiles (Supplemental File 1). Gut prevalence had a strong phylogenetic signal (Pagel's $\lambda = 0.97$, likelihood-ratio $p < 10^{-22}$), meaning that it was strongly correlated with the evolutionary relatedness of species. This emphasizes the need for phylogeny aware modeling so that signal linking genes to prevalence will not be drowned out by shared variation in gene content between closely-related species.

To demonstrate the effect of phylogenetic correlation empirically, we fit both a standard linear model and a phylogenetic linear model for each of the four common gut phyla and all genes present in that phylum. These models relate the logit-transformed prevalence of different species in the phylum to a gene's presence/absence in their pangenomes. The models have the following form:

$$\text{logit}\, P(T) \equiv \log\left(\frac{P(T)}{1-P(T)}\right) = \beta_0 + \beta_{1,g} g_T + \epsilon_T \tag{2}$$

where $P(T)$ is the prevalence of species T, $\beta_{1,g}$ captures the association of gene g with $P(T)$ across species, and the other terms are defined above. The unit of measurement is a species, and each species has a value for $P(T)$ and g_T. Recall that ϵ_T is the residual variation in logit-prevalence, which is independent and normally distributed in the standard linear model but has a distribution encoding correlations proportional to species relatedness in the phylogenetic linear model (see Methods). For both standard and phylogenetic linear models, separate models were fit for each phylum. We modeled associations for 144,651 genes total across the four phyla, fitting 381,846 models total (since some genes are present in multiple phyla).

We used the parameter estimates and their standard errors from fitted models to test null hypotheses of the form $H_0 : \beta_{1,g} = 0$, meaning gene g is not associated with gut prevalence of species in a particular phylum. The p-values were adjusted for multiple testing using the false discovery rate (FDR)(see Methods). We found 9,830 genes positively associated with logit-prevalence within at least one phylum (FDR $q \leq 0.05$) using phylogenetic linear models. We observed that 75% of the significant genes from these tests had effect sizes larger than (Bacteroidetes) 0.93, (Firmicutes) 1.03, (Proteobacteria) 0.35, and (Actinobacteria) 2.04, which are within the range of effect sizes for which phylogenetic linear models showed good performance in simulations (see above).

With standard linear models our tests identified 25,185 genes associated with gut prevalence, substantially more than with phylogenetic linear models (17.4% versus 6.8% of total) and, based on our simulations, likely including many false positives. The top results of phylogenetic versus standard linear models (Figure 2) illustrate the pitfalls of not correcting for phylogenetic correlation. Using the standard model, we recover associations such as those seen in Figure 2A-B: a subunit of dihydroorotate dehydrogenase in Bacteroidetes (Figure 2B) and in Firmicutes, a particular type of glutamine synthetase (Figure 2A). While these associations might look reasonable at a first glance, on closer inspection, they depend on the fact that these genes are near-uniformly present in entire clades of bacteria. These clades are, in general, more prevalent in the gut compared to the rest of the species in the tree. However, any finer structure relating to differences between close neighbors is lost.

While this alone does not necessarily constitute evidence *against* these genes having adaptive functions in the gut, we do expect that matched pheno- and genotypic differences between close phylogenetic neighbors offer stronger evidence for an association. An analogy can be drawn with genome-wide association mapping in humans: models that do not account for correlations between sites caused by population structure, as opposed to selective pressure, will tend to identify more spurious associations. In contrast, because the

phylogenetic null model "expects" trait correlations to scale with the evolutionary distance between species, this approach will tend to upweight cases where phylogenetically close relatives have different phenotypes and where distant relatives have similar phenotypes. This leads to the identification of candidate genes that capture more variation between close neighbors (Figure 2C-D). Thus, phylogenetic linear models will identify genes whose presence in genomes is more frequently changing between sister taxa in association with a trait.

We provided further evidence that this trend is true in general by calculating the phylogenetic signal of the top hits from each model using Ives and Garland's α [32]. This statistic captures the rate of transitions between having and not having a binary trait (here, a gene) across a tree; higher values therefore correspond to more disagreement between closely related species and lower values correspond to more agreement. Indeed, across all four phyla, the linear model identified gene families with significantly lower Ives-Garland α than the phylogenetic model (Figure 2E, linear model $p < 10^{-16}$), indicating that these genes' presence versus absence tended to be driven more by clade-to-clade differences (i.e., shared evolution).

These results show that standard linear models will identify genes that are truly important for colonizing an environment, such as the healthy human gut, as well as other genes unrelated to the environment but also common in clades with many species that are present in the environment. The latter set will likely include many false positive associations from the perspective of understanding functions necessary for living in the environment. Phylogenetic linear models overcome this problem by accounting for correlations due to both phenotypes and genotypes being more similar amongst closely related species. These conclusions are supported by our simulations and by an *in vivo* functional screen (see below).

2.4 Gene families associated with gut prevalence provide insight into colonization biology

Several of the gene families that we observe to be associated with gut prevalence have previously been linked to gut colonization efficiency. For example, in Firmicutes, we noticed that several top hits were annotated as sporulation proteins (e.g., "stage IV sporulation protein B", FIG00004463, Figure 1C). Sporulation is known to be a strategy for surviving harsh environments (such as acid, alcohol, and oxygen exposure) that is used by many, but not all, members of Firmicutes. Resistance to oxygen (aerotolerance) is particularly important because many gut Firmicutes are strict anaerobes [33], sporulation is known to be an important mechanism of transmission and survival in the environment (reviewed in Swick et al. [34]), and sporulation ability has been linked to transmission patterns of gut microbes [29]. Our result associating sporulation proteins to gut prevalence provides further evidence for sporulation as a strategy that is generally important for the propagation and fitness of gut microbes.

In Bacteroidetes, we observed an association between gut prevalence and the presence of a pair of gene families putatively assigned to the GAD operon, namely, the glutamate decarboxylase *gadB* and the glutamate/gamma-aminobutyric acid (GABA) antiporter *gadC*. These genes show a complex pattern of presence that is strongly correlated with gut prevalence (Figure 2D). Results from research in Proteobacteria, where these genes were first described, shows that their products participate in acid tolerance. L-glutamate must be protonated in order to be decarboxylated to GABA; export of GABA coupled to import of fresh L-glutamate therefore allows the net export of protons, raising intracellular pH [35]. It was previously hypothesized that this acid tolerance mechanism allowed bacteria to survive the harshly acidic conditions in the stomach: indeed, if disrupted in the pathogen *Edwardsiella tarda*, gut colonization in a fish model is impaired [36]. *Listeria monocytogenes* with disrupted Gad systems also become sensitive to porcine gastric fluid [37]. However, while it has previously been shown that gut *Bacteroides* do contain homologs for at least one of these genes [35], their functional importance has not yet been demonstrated in this phylum. Our results provide preliminary evidence that this system may be important in Bacteroidetes as well as in Proteobacteria.

2.5 Using body sites as a control allows us to differentiate general dispersal from a specific gut advantage

The previous analyses have focused on modeling the phenotype of overall prevalence in the human gut. However, microbes could be prevalent in the gut for at least two main reasons. First, they could be specifically well-adapted to the human gut; second, they could simply be very common in the environment (i.e., highly

dispersed). The presence or absence of a gene family could enhance either of these properties. Some genes might, for example, confer improved stress tolerance that was adaptive across a range of harsh conditions, while others might allow, for example, uptake and catabolism of metabolic substrates that were more common in the human gut than in other environments.

With this in mind, we analyzed the relative enrichment of microbes in the gut over other human body sites in 127 individuals from the Human Microbiome Project study [38]. We chose other body sites as a control because the physical distance between sites within a host is much smaller than the distance between people, and microbes from one body site are likely to be commonly, if transiently, introduced to other body sites (e.g., skin to oral cavity). To find specifically gut-associated genes, we used the phylogenetic linear model to regress gene presence/absence on the logit-transformed conditional probability $P(B = gut|T)$, i.e., the probability that a body site B was the human gut given that a particular species T was observed, which we estimated using Laplace regularization (see Methods). We identified 4,672 genes whose presence in bacterial genomes was associated with those species being present in the gut versus other body sites in at least one phylum (397 in Bacteroidetes, 1,572 in Firmicutes, 1,284 in Proteobacteria, and 1,507 in Actinobacteria).

Overall, the effect sizes for genes learned from this body site-specific model correlated only moderately with those learned from the "gut prevalence" models (median $R^2 = 0.06$, range -0.06—0.24), indicating that these two quantitative phenotypes describe distinct phenomena. Additionally, the overlap between significant ($q \leq 0.05$) hits for both models was small (median Jaccard index 0.054, range 0.011—0.089). These results are not surprising given that our regularized estimates of gut specificity were only moderately correlated with overall gut prevalence (Spearman's $\rho = 0.33$, Supplemental Figure S1). This may arise from different genes being involved in dispersal or adaptation to many different environments versus those involved in adaptation specifically to the gut.

Indeed, when we compare enrichments for genes significant in either the body site or overall prevalence models alone (i.e., genes with $q \leq 0.05$ in one model but $q > 0.5$ and/or wrong sign of effect size in the other), we observe large functional shifts (Figure 3). For example, in the gut prevalence model, but not the body site-specific model, Firmicutes were strongly enriched for "dormancy and sporulation" ($q = 8.7 \times 10^{-7}$). Because sporulation is likely useful in a wide range of environments beyond the gut, this result seems intuitive. Body site-specific results for Firmicutes were instead enriched for genes involved in "phosphate metabolism" ($q = 0.12$) and in particular the term "high affinity phosphate transporter and control of PHO regulon" ($q = 0.05$).

We also observed biologically-justified individual gene families that were significant in the body site-specific model but not the overall gut prevalence model. In Firmicutes, for example, carnitine dehydratase and bile acid 7-alpha dehydratase were both significant only in the body site-specific model, suggesting a specific role for these genes within the gut environment. Indeed, bile acids are metabolites of cholesterol that are produced by vertebrates and thus unlikely to be encountered outside of the host. While the metabolite L-carnitine is made and used in organisms spanning the tree of life, it is particularly concentrated in animal tissue and especially red meat, and cannot be further catabolized by humans [39], making it available to intestinal microbes. Bile acid transformation by gut commensals is a well-established function of the gut microbiome, with complex influences on health (reviewed in Staley et al. [40]).

In Bacteroidetes, we found that a homolog of the autoinducer 2 aldolase *lsrF* was significant only in the body site-specific model. Autoinducer 2 is a small signaling molecule produced by a wide range of bacteria that is involved in interspecies quorum sensing. The protein *lsrF*, specifically, is part of an operon whose function in *Escherichia coli* is to "quench" or destroy the AI-2 signal [41]. Further, an increase of the AI-2 signal has been shown to decrease the Bacteroidetes/Firmicutes ratio *in vivo* in the intestines of streptomycin-treated mice [42]. Degrading this molecule is therefore a plausible gut-specific colonization strategy for gut Bacteroidetes. These discovered associations make the genes involved, including many genes without known functions or roles in gut biology, excellent candidates for understanding how bacteria adapt to the gut environment.

2.6 Deletion of gut-specific genes lowers fitness in the mouse microbiome

Beyond finding evidence for the plausibility of individual genes based on the literature, we were interested in whether more high-throughput experimental evidence supported the associations we found between gut colonization and gene presence. To interrogate this, we used results from an *in vivo* transposon-insertion

screen of four strains of *Bacteroides*. This screen identified many genes whose disruption caused a competitive disadvantage in gnotobiotic mice, as revealed by time-course high-throughput sequencing; 79 gene families significantly affected microbial fitness across all four strains tested [43]. Determining agreement with this screen is somewhat complicated by the fact that we associated gene presence to gut specificity across all members of the phylum Bacteroidetes, and not only within the *Bacteroides* genus. Significance of overlap therefore depends on what we take as the null "background" set, the cutoff used for significance, and the set of results from the screen we choose as true positives (Supplementary Table S2).

Despite these complications, this analysis clearly showed that the 79 genes whose disruption led to lower fitness in the murine gut across all four *Bacteroides* species were over-represented among our predictions for gut-specific genes (odds ratio = 4.39, $q = 8.3 \times 10^{-3}$), and remained so if we only considered the gene families that were present in all *Bacteroides* species (odds ratio= 7.02, $q = 3.0 \times 10^{-3}$) (Table 1). Interestingly, we observed the opposite pattern for the overall prevalence model: the prevalence-associated genes we identified were actually depleted for genes found to be important *in vivo* (odds ratio = 0.18, $q = 7.7 \times 10^{-3}$). We believe that this is because the body-site-specific model, like the experiment, focused specifically on colonization efficiency, while the overall gut prevalence model would have included genes involved in persistence and dispersal in the environment and transfer between hosts. This experimental evidence supports the idea that environment-specific phylogenetic linear models truly identify genes that are important for bacteria to colonize an environment.

2.7 Proteobacterial gene families are associated with microbes more prevalent in Crohn's disease

The above analyses were performed with respect to the gut of healthy individuals from the mainly post-industrial populations of North America, Europe and China. However, we also know that taxonomic shifts are common between healthy guts versus the guts of individuals from the same population with diseases such as type 2 diabetes, colorectal cancer, rheumatoid arthritis, and inflammatory bowel disease (reviewed in Wang et al. [44]). One explanation for these results is that sick hosts select for specific microbial taxa, as with the links previously observed between Proteobacteria and the inflammation that accompanies many disease states [45]. Since gut microbes have also been implicated in altering disease progression (reviewed in Lynch and Pedersen [46]), identifying genes associated with colonizing diseased individuals may afford us new opportunities for intervention.

To identify microbiome functions that could be involved in disease-specific adaptation to the gut, we looked for genes that were present more often in microbes that discriminated case from control subjects. Specifically, we compared $n = 38$ healthy controls from the MetaHIT consortium to $n = 13$ individuals with Crohn's disease [47, 48]. Similar to our analysis of gut versus other body sites, we used the conditional probability that a subject had Crohn's disease *given* that we observed a particular microbe in their gut microbiome $P(CD|T)$ (see Methods). We identified 1,904 genes whose presence in bacterial genomes is associated with Crohn's after correcting for phylogenetic relationships in at least one phylum (800 in Bacteroidetes, 272 in Firmicutes, 529 in Proteobacteria, and 319 in Actinobacteria).

Three of our top Proteobacterial associations were annotated as fimbrial proteins, including one predicted to be involved specifically in the regulation of type 1 fimbriae, or pili (FimE, association $q = 4.0 \times 10^{-6}$), cell surface structures involved in attachment and invasion. Crohn's pathology has been linked to an immune response to invasive bacteria, and adherent-invasive *E. coli* (AIEC) appear to be overrepresented in ileal Crohn's [49]. In an AIEC *E. coli* strain isolated from the ileum of a Crohn's patient, type 1 pili were required for this adherent-invasive phenotype [50]. Chronic infection by AIEC strains was also observed to lead to chronic inflammation, and to an increase in Th17 cells and a decrease in $CD8^+$ T cells similar to that observed in Crohn's patients [51].

An additional striking feature of the results was the number of Proteobacterial proteins associated with greater risk of Crohn's that were annotated as being involved in the the type III, IV, VI, and ESAT secretion systems (Fisher's test $q = 0.13$). On further investigation, we found that these proteins were actually all predicted to be involved in conjugative transfer, a process by which gram-negative bacteria in direct physical contact share genetic material. More specifically, many of these genes were homologs of those involved in an "F-type" conjugal system for transferring IncF plasmids, which can be classified as a variety of type IV secretion system [52]. Previously, in a mouse model, gut inflammation was shown to stimulate efficient

horizontal gene transfer in Proteobacteria by promoting blooms of *Enterobacteriaceae* and thus facilitating cell-to-cell contact [53]. Future work will be required to determine whether this increased conjugation is a neutral consequence of inflammation, a causative factor, or provides a selective advantage in the inflamed gut.

3 Discussion

The present analyses represent a first look into what can be learned by combining shotgun metagenomics with phylogenetically-aware models. Several extensions to our work could be made in the future. First, in addition to modeling prevalence, for instance, we could model abundance using a phylogenetic linear model with random effects [54], potentially allowing us to learn what controls the steady-state abundance of species in the gut. Additionally, we could also use these models to screen for epistatic interactions, which would be near-intractable even in systems with well-characterized genetic tools, but for which a subset of hypotheses could be validated by, e.g., comparing the fitness of wild-type microbes with double knockouts. While controlling the total number of tests would still be important to preserve power, an automated, computational approach to detecting gene interactions would still offer important savings in time and expense over developing a genome-wide experimental library of multiple knockouts per organism under investigation.

Currently, these analyses estimate species abundance and gene presence/absence from available sequenced isolate genomes. However, it has been estimated that on average 51% of genomes in the gut are from novel species [29]. Especially for case/control comparisons, using information from metagenomic assemblies could enable quantification of species with no sequenced representatives, and would yield a more accurate estimate of the complement of genes in the pangenome for species that do have sequenced representatives. This would be particularly helpful in gut communities from individuals in non-industrialized societies that are enriched for novel microbial species [29]. In fact, genes then could be treated as quantitative variables (e.g., coverage or prevalence) rather than binary, which is possible for covariates in phylogenetic linear models and simply changes the interpretation of the association coefficient $\beta_{1,g}$.

Another potential extension would be to model prevalence and environment-specific prevalence for taxa other than the species clusters analyzed in this study. We focused on four prevalent and abundant phyla of bacteria, but our methods could be applied more broadly as long as quantitative traits and genotypes could be accurately estimated. Phylogenetic linear modeling could also be applied directly to genera or higher taxonomic groups, although both traits and genotypes would be averages over more diverse sets of genomes, which could result in associations with different signs canceling out. As more genome and metagenome data is generated for microbial populations over time, extensions of phylogenetic linear modeling (e.g., with random effects [54]) may also be useful for studying associations between traits and evolving gene copy number and single nucleotide variants at the strain level. This application would require accurate trees with strains as leaves, each with estimates of a trait and genotype. Beyond prevalence, other traits will also be interesting to investigate, especially experimentally measured phenotypes from high throughput screens and other techniques that complement genomics.

In summary, using phylogenetic linear models, we were able to discover thousands of specific gene families associated with quantitative phenotypes calculated directly from data: overall gut prevalence, a specificity score for the gut over other body sites, and a specificity score for the gut in Crohn's disease versus health. Importantly, we have shown through simulation and real examples that standard linear models are inadequate for this task because of an unacceptably high false-positive rate under realistic conditions. Furthermore, many of the results we found also have biological plausibility, both from the literature on specific microbial pathways and from a high-throughput *in vivo* screen directly measuring colonization efficiency. In addition to these expected discoveries, we also found thousands of novel candidates for understanding and potentially manipulating gut colonization. These results illustrate the potential of integrating phylogeny with shotgun metagenomic data to deepen our understanding of the factors determining which microbes come to constitute our gut microbiota in health and disease.

4 Methods

4.1 Species definition

We utilized the previously published clustering of 31,007 high-quality bacterial genomes into 5,952 species from the MIDAS 1.0 database [29] (`http://lighthouse.ucsf.edu/MIDAS/midas_db_v1.0.tar.gz`). These species clusters are sets of genomes with high pairwise sequence similarity across a panel of 30 universal, single-copy genes. The genomes in each species clustering have approximately 95% average genome-wide nucleotide identity, a common "gold-standard" definition of bacterial and archaeal species [55]. These species-level taxonomic units are similar to, but can differ from, operational taxonomic units (OTUs) defined solely on the basis of the 16S rRNA gene.

Taxonomic annotations for each species were drawn from the MIDAS 1.0 database. Some taxonomic annotations of species in the MIDAS database were incomplete; these were fixed by searching the NCBI Taxonomy database using their web API via the `rentrez` package [56] and retrieving the full set of taxonomic annotations.

4.2 Pangenomes

Pangenomes for all species used in this study were downloaded from the MIDAS 1.0 database. As previously described [29], pangenomes were constructed by clustering the DNA sequences of the genes found across all strains of each species at 95% sequence identity using UCLUST [57]. Pangenomes were functionally annotated based on the FIGfams [58] which were included in the MIDAS databases and originally obtained from the PATRIC [59] database. Thus, each pangenome represents the set of known, non-redundant genes from each bacterial species with at least one sequenced isolate.

4.3 Phylogenetic tree construction

The tree used for phylogenetic analyses was based on the tree from Nayfach et al. [29] based on an approximate maximum likelihood using FastTree 2 [60] on a concatenated alignment (using MUSCLE [61]) of thirty universal genes. Thus, each tip in the tree represents the phylogenetic placement for one bacterial species. For the current analyses, the tree was rooted using the cyanobacterium *Prochlorococcus marinus* as an outgroup, and the tree was then divided by phylum, retaining the four most prevalent phyla in the human gut (Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria). One Actinobacterial species cluster, the radiation-resistant bacterium *Kineococcus radiotolerans* SRS30126, was dropped from the tree because it had an extremely long branch length, indicating an unusual degree of divergence. Finally, phylum-specific trees were made ultrametric using the `chronos` function in the R package `ape` [62], assuming the default "correlated rates" model of substitution rate variation.

4.4 Estimating species abundance across human associated metagenomes

Metagenome samples were drawn from healthy subjects in the Human Microbiome Project [38], the MetaHIT consortium [47, 48], a study of glucose control [63], and a study of type 2 diabetes [64]. Accession numbers were identified using the aid of SRAdb [65] and downloaded from the Sequence Read Archive (SRA) [66]. The relative abundance of bacterial species in the metagenomes was estimated using MIDAS v1.0 [29], which maps reads to a panel of 15 phylogenetic marker genes.

Accession IDs used can be found in Supplementary Table S3.

4.5 Modeling overall gut prevalence and environmental specificity scores

Prevalence can be understood as the probability of observing a particular microbe in a given environment e, $P(T|E = e)$. (We use the term *overall prevalence* to refer specifically to the prevalence in the healthy gut in this study.) Modeling prevalence as a probability provides an intuitive justification of the logit transform, which is used for the response variable in logistic and binomial regression. Because we integrated data from multiple studies (see above) and did not want one study to dominate the results, instead of simply using

the counts across all samples to determine the prevalence, we used weighted counts $W(T, E = e)$, where the weights came from the number of samples per study $s \in S$:

$$W(T, E = e) = \left(\sum_{s \in S} \#\left(E = e, S = s\right) \right) \left(\sum_{s \in S} \frac{\#(T, E = e, S = s)}{\#(E = e, S = s) \cdot |S|} \right)$$

In order to avoid probabilities of 0 or 1, probabilities were based on prevalence with one additional "present" and one additional "absent" pseudocount; this procedure is equivalent to a maximum *a posteriori* estimate derived from placing a uniform Beta$(1, 1)$ prior on a binomially-distributed parameter:

$$P(T|E = e) = \frac{W(T, E = e) + 1}{\#(E = e) + 2}$$

Beyond prevalence, we were also interested in *environmental specificity*, or the discriminative power of a given microbe among a given set of environments, such as body sites or disease states. Statistically, we define the environmental specificity score to be the conditional probability $P(E = e|T)$ that, given that a particular species T is observed in a sample, that the sample is from environment $e \in E$. By Bayes' rule, we can see that:

$$P(E = e|T) = \frac{P(T|E = e)P(E = e)}{P(T)} \equiv \frac{P(T|E = e)P(E = e)}{\sum_e P(T|E = e)P(E = e)}, e \in E \tag{3}$$

This conditional probability can be estimated directly from the environment-specific prevalences $P(T|E = e)$ and the prior probabilities $P(E = e)$ of a sample being taken from a given environment. However, if we simply plug in the maximum-likelihood estimates of $P(T|E = e)$ to the above equation, infrequent observation of a species (for example, $0/38$ in healthy subjects versus $1/13$ in subjects with a disease) will yield inappropriately extreme estimates of $P(E|T)$. We therefore use a regularized estimate of $P(E = e|T)$ that incorporates a prior centered on $P(E = e)$.

More specifically, we use a maximum a posteriori (MAP) estimator of $P(E = e|T)$ with a Laplace prior centered on $\text{logit}(P(E = e))$. Laplace regularization is used in the Bayesian Lasso to perform L1-norm penalization of parameter estimates. We use it in a similar way, i.e., to reduce the variance in $P(E = e|T)$ by shrinking more uncertain estimates towards the prior, and to reduce the total number of species for which $P(E = e|T) \neq P(E = e)$. The general form of MAP estimates is the following:

$$p_{MAP}(D, Q) = \text{argmax}_p \mathscr{L}(D|p)\mathscr{L}_\pi(D|p; Q) \tag{4}$$

where p is the parameter being estimated, D represents the data (or sufficient statistics derived from it), Q represents the set of hyperparameters (parameters of the prior distribution), $\mathscr{L}$ represents the likelihood function of the distribution from which the data is assumed to be drawn, and $\mathscr{L}_\pi$ represents the likelihood of the prior distribution (without which the estimator reduces to the maximum-likelihood estimator). In our case, we model the data as having the following distribution:

$$\begin{aligned} D &\sim Binomial(x, n) \\ x &= \frac{pt}{q} \\ \text{logit}(p) &\sim Laplace(q, b) \end{aligned}$$

The parameter being estimated is p, corresponding to $P(E = e|T)$. We observe n, the number of samples from environment e in the dataset (as well as k, the number of samples in environment e in which species T is observed), and t, the weighted prevalence across classes $P(T) = \sum_e P(T|E = e)P(E = e)$ (here, $P(T|E = e)$ are maximum-likelihood estimates directly from the data). The hyperparameters of this model are b, a tuning hyperparameter corresponding to the strength of regularization, and q, the prior probability $P(E = e)$. In the above formulation, we use x to represent the prevalence $P(T|E = e)$; its value $\frac{pt}{q} \equiv \frac{P(E=e|T)P(T)}{P(E=e)}$

follows from Bayes' rule. The MAP estimate of $p = P(E = e|T)$ is therefore obtained through the following maximization:

$$\widehat{p}_{MAP}(\{n,k,t\},\{q,b\}) = \text{argmax}_p \left[\left(\binom{n}{k} \left(\frac{pt}{q}\right)^k \left(1 - \frac{pt}{q}\right)^{n-k} \right) \left(\frac{1}{2b} \exp(-\frac{|\text{logit}(p) - \text{logit}(q))}{b} \right) \right] \quad (5)$$

To choose appropriate, dataset-specific values of b, which controls how much estimates of $P(E = e|T)$ are shrunk back to the prior, we performed simulations based on these datasets, in which presence vs. absence $P_{t,e}$ of species $t \in T$ across environments $e \in E$ were modeled as follows:

$$\begin{aligned} P_{t,e} &\sim Binomial(x_{t,e}, n_e) \\ x_{t,e} &= \begin{cases} y_t & (e \neq e_1) \\ \text{logistic}(\text{logit}(y_t + F_t)) & (e = e_1) \end{cases} \\ F_t &\sim \begin{cases} f \cdot (2 \cdot (Bernoulli(g)) - 1) & (t \notin T_0) \\ 0 & (t \in T_0) \end{cases} \\ y_t &\sim Beta(a, b) \end{aligned}$$

In other words, for each species t in different environments e, presence-absence $P_{t,e}$ was modeled as a binomial random variable. The success parameter from this binomial was drawn from a Beta distribution with parameters a and b, which were fit from a single environment in the corresponding real dataset using maximum-likelihood, thus ensuring that the simulated species had similar baseline prevalences as real species. In species with no difference between environments $t \in T_0$, the true prevalence x_t was set to be equal between the environment of interest e_1 and all other environments; in species with true differences between environments ($t \notin T_0$), in contrast, the effect size f was either added or subtracted from the logit-prevalence (with the parameter g controlling the proportion of positive true effects). The number of null species $||T_0||$ was set to 25% of the total number of simulated species $||T||$, which was matched to the real dataset.

For a given simulated dataset and value of b, the false positive rate (FPR_b) and the true positive rates for $F > 0$ and $F < 0$ (TPR_{pos} and TPR_{neg}, respectively) were calculated:

$$FPR_b = \#(|P(E = e|t \in T_0) - P(E = e)| > \epsilon)/||T_0||$$

$$TPR_{pos_b} = \#(P(E = e|F_t > 0) - P(E = e) > \epsilon)/\#(F_t > 0)$$

$$TPR_{neg_b} = \#(P(E = e) - P(E = e|F_t < 0) > \epsilon)/\#(F_t < 0)$$

ϵ is a tolerance parameter set at $P(E = e) \cdot 0.005$ to account for numerical error. The tuning parameter b was then optimized according to the following piecewise continuous function, which increases from 0 to 1 until the false positive rate drops to 0.05 or lower (in order to guide the optimizer), and then increases above 1 in proportion to the average (geometric mean) of the positive and negative effect true positive rates:

$$\text{argmax}_b \begin{cases} 1 - FPR_b & FPR_b > 0.05 \\ 1 + \sqrt{TPR_{pos_b} \times TPR_{neg_b}} & FPR_b \leq 0.05 \end{cases}$$

Given $f = 2$, for Crohn's disease, b was estimated at $b = 0.14$ and for the body site specificity, b was estimated at $b = 0.19$. (Changing f to 1 or 0.5, or changing g to 0.1 or 0.9, resulted in very similar estimates of b.)

Calculating environmental specificities requires a prior $P(E = e)$. In the case of the environmental specificity for Crohn's disease, this prior was from epidemiological data [67] and fixed at 0.002. In the case of body site specificity, we instead assumed an uninformative (i.e., uniform) prior across the body sites considered (supragingival plaque, subgingival plaque, buccal mucosa, posterior fornix, tongue dorsum, anterior nares, retroauricular crease, and gut, from the Human Microbiome Project [38]), meaning that in the absence of other information, all body sites were considered to be equally likely. To avoid introducing study effects, when computing the disease-specific conditional probabilities, we used only Crohn's cases and

controls from the same MetaHIT cohort ([47, 68]), and when computing body site specificities, we used only samples from the Human Microbiome Project [38].

This process can be visualized in Supplemental Figure S2. As an example, two species, one with very little predictive power for Crohn's disease (*Bacillus subtilis*) and another with high predictive power (*Bacteroides fragilis*), are compared. Without regularization, *Bacillus subtilis* actually appears to be a better predictor. This is because with few observations in each condition, estimating the predictiveness of *B. subtilis* for Crohn's, $P(E = e|T)$, is noisy. *B. subtilis* appears in 1 out of 13 Crohn's samples (0.077) versus 1 out of 38 (0.026) healthy samples, while *B. fragilis* appears in 13 out of 13 Crohn's samples (1.00) but also 24 out of 38 healthy samples (0.63) (Supplemental Figure S2a).

Unregularized, $P(E = e|T) = P(T|E = e)P(E = e)/P(T)$. Since the prior for Crohn's $P(E = e)$ is low, $P(E = e|T) \approx P(T|E = e)P(E = e)/P(T|E \neq e)$, meaning that the estimate is mainly determined by the ratio of prevalences in each environment. Naively, *B. subtilis* is 2.9 times more likely to appear in Crohn's vs. normal, while for *B. fragilis* that ratio is only 1.6. However, intuitively, we expect that the estimate for *B. fragilis* would be much more stable to perturbation: removing one Crohn's observation of *B. fragilis* would only drop the ratio from 1.6 to 1.5, while removing the lone Crohn's observation for *B. subtilis* takes the ratio from 2.9 to zero. This difference in how confidently $P(E = e|T)$ is estimated can be seen quantitatively by comparing the likelihood distributions for the unregularized estimates of $\text{logit}(P(E|T))$ given the data (Supplemental Figure S2C). The distribution for *Bacillus subtilus* is flatter (note y-axis) and more spread out than the distribution for *Bacteroides fragilis*. Therefore, when we regularize using the Laplace prior (Supplemental Figure S2d), the MAP estimate for *B. subtilis,* but not for *B. fragilis,* is dominated by the peak at $P(E = e)$, i.e., 0.002 (Supplemental Figure S2e). The tuning parameter b (discussed above) controls the width of the Laplace prior.

4.6 Phylogenetic and non-phylogenetic linear models

Throughout the text, we used phylogenetic and non-phylogenetic linear models to identify FIGfam gene families associated with presence of bacterial species in the human gut. The response variables in the models we fit were logit-transformed probabilities. These probabilities fell into two categories: 1. the overall probability of observing a particular species T in the gut $P(T)$ (i.e., prevalence), and 2. the conditional probabilit y of a sample/subject coming from some particular state $P(\bullet|T)$ (specifically, a given body site or a given disease state) given an observation of this microbe.

Per-gene-family models were fit according to the following simple model. For each gene family (in this case, FIGfams, a grouping of orthologous genes [58]) in each phylum, we fit a model of the following form:

$$\log\left(\frac{P(T)}{1 - P(T)}\right) = \beta_0 + \beta_1 g_t + \epsilon_t \tag{6}$$

where $P(T)$ is the prevalence of a species (equivalently, the estimated probability of observing that species), β are model parameters, g_t is an indicator variable that is 1 if species t has gene g in its sequenced pan-genome and 0 otherwise, and ϵ_t are the residuals, i.e., the unmodeled variation in the species prevalence.

The phylogenetic and standard linear models are very similar, except for the assumptions about the distribution of the residuals. In the standard linear model, the residuals are taken to be independently and identically distributed as a normal distribution, i.e., $\epsilon_t \underset{iid}{\sim} N(0, \sigma^2)$. In the phylogenetic model, in contrast, the residuals are not independent: rather, they are correlated based on the phylogenetic relatedness of the species. They are therefore distributed $\epsilon_t \sim MVN(0, \Sigma)$ with the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma^2 & & \\ \vdots & & \ddots & \\ \sigma_{n,1} & & & \sigma^2 \end{bmatrix}$$

where n is the number of species, σ^2 is the overall variance, and $\sigma_{1,2}$ is the shared variance between species 1 and species 2. Under the assumption of the phylogenetic model, this shared variance is proportional to the distance between the last common ancestor of species 1 and 2 and the root of the tree; very closely-related species have a common ancestor that is far from the root, while the last common ancestor of two unrelated

species is the root node itself. This method was first described in Grafen [25]; for this study, we use the implementation in the phylolm R package [69].

β_1 parameters were tested for a significant difference from 0 and the resulting p-values were converted to q-values using Storey and Tibshirani's FDR correction procedure [70, 71].

To find genes associated specifically with prevalence in the gut as opposed to other body sites, or genes specifically associated with prevalence in a particular disease state, we again used phylogenetic linear models, this time using the logit-transform of the appropriate conditional probability:

$$\log\left(\frac{P(\bullet|T)}{1-P(\bullet|T)}\right) = \beta_0 + \beta_1 g_t + \epsilon_t \tag{7}$$

4.7 Enrichment analysis

Enrichment analysis was performed using SEED subsystem annotations for FIGfams [72, 58]. Each subsystem was tested for a significant overlap with significant hits from the linear models ($q \leq 0.05$), given the set of FIGfams tested, by Fisher's exact test. The p-values were corrected using the Benjamini-Hochberg procedure [73] and an FDR of 25% was set for detecting significant enrichments.

4.8 Overlap with *in vivo* results

Results of the screen were obtained from the Supplemental Material of Wu et al. (downloaded on 2017 May 3) [43]. We used the genes the authors identified as having significant effects in both diet conditions in all species. Genes were mapped to FIGfams by matching identifiers in the Supplemental Material to genome annotations from PATRIC [59]. Significance of overlap between these genes and the results for the Bacteroidetes phylum from the body-site-specific or overall models was determined by Fisher's exact test.

4.9 Codebase

The code used to perform these analysis is available at `http://www.bitbucket.com/pbradz/plr` in the form of an Rmarkdown notebook.

5 Author contributions

- Patrick H. Bradley. Roles: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing
- Stephen Nayfach. Roles: Data Curation, Software, Resources, Writing – Review & Editing
- Katherine S. Pollard. Roles: Conceptualization, Methodology, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

6 Funding statement

Funding for this research was provided by NSF grants DMS-1069303 and DMS-1563159, Gordon & Betty Moore Foundation grant #3300, and institutional funds from the Gladstone Institutes. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7 Acknowledgements

The authors would like to thank Joshua Ladau, Nandita Garud, and other members of the Pollard and Turnbaugh groups, as well as attendees of the 2017 Keystone meeting on the Microbiome in Health and Disease and attendees of the Second Workshop in Statistics and Algorithmic Challenges in Microbiome Data Analysis (SACMDA2), for helpful discussions.

Figure legends

Figure 1: **Failing to account for tree structure results in an elevated false positive rate.** Continuous and binary traits were simulated across the trees for the four phyla under consideration. A-B show results for the null of no true correlation between the continuous and binary traits. A) Histogram of p-values for simulated traits on the Bacteroidetes tree, using phylogenetic (left) or standard (right) linear models. The phylogenetic model distribution was similar to a uniform distribution, while the standard model was very anticonservative, having an excess of small p-values. B) False positive rate (Type I error rate) at $p = 0.05$ for the phylogenetic and standard models. C) Traits with varying levels of "true" association spanning values we observed in real data were simulated, and power was computed using phylogenetic linear models.

Figure 2: **Examples of hits from standard linear (blue highlights) and phylogenetic (orange highlights) models.** In each panel, the tree on the left is colored by species prevalence (black to orange), while the tree on the right is colored by gene presence/absence (blue to black). Selected species are displayed in the middle; lines link species with the leaves to which they refer. The color of the line matches the color of the leaf. A-B) The standard model recovered hits that matched large clades but without recapitulating fine structure. C-D) The phylogenetic model recovered associations for which more of the fine structure was mirrored between the left-hand and right-hand trees, as exemplified by the species labeled in the middle. E) Violin plots of Ives-Garland α, a summary of the rate of gain and loss of a binary trait across a tree, for genes significantly associated with prevalence in the standard (left, blue) and phylogenetic (right, orange) linear models. Horizontal lines mark the median of the distributions. The phylogenetic (orange) and standard linear (blue) models were significantly different for each phylum (Wilcox test for Bacteroidetes: 4×10^{-6}; Firmicutes: 7×10^{-11}; Proteobacteria: 2×10^{-22}; Actinobacteria: 2×10^{-22}).

Figure 3: **Comparison of results from the overall prevalence and body-site specific models for Firmicutes.** FDR-corrected significance (as $-\log_{10}(q)$) of the overall model is plotted on the horizontal axis, whereas the same quantity for the body-site-specific model is plotted on the vertical axis. All FIGfams significant ($q \leq 0.05$) in at least one of the two models are plotted as contour lines: FIGfams significant in the overall prevalence model (and possibly also the gut specific model) are plotted in orange, while FIGfams significant in the gut specific model (and possibly also the overall prevalence model) are plotted in blue. Selected SEED subsystems are displayed as colored points (legend), and selected individual genes are plotted as black points.

Figure 4: **Genes involved in conjugative transfer are associated with Crohn's disease-enriched species.** The conjugation transcriptional regulator traR is plotted as an example. The left-hand tree is colored by each species' disease specificity score, i.e., the conditional probability of Crohn's given the observation of a given species (grey, which represents the prior, to orange, which represents a higher conditional probability). The right-hand tree is colored by gene presence-absence (grey, meaning absent, or blue, meaning present). The mirrored patterns drive the phylogeny-corrected correlation.

Table 1: **Assessment of agreement between the *in vivo* results from Wu et al. [43] and gut-specific ("bodysite") vs. gut prevalence ("overall") phylogenetic models.** The background sets for enrichment tests were defined as follows: "all tested" (all gene families for which a phylogenetic model was fit), "Bacteroides (core or variable)" (all gene families with at least one representative in *Bacteroides* genome cluster pangenomes), "Bacteroides (core only)" (gene families that were present in all *Bacteroides* genome cluster pangenomes), "Bacteroides (variable only)" (gene families present in some but not all *Bacteroides* genomes clusters), and "Bacteroides thetaiotaomicron only" (only gene families present in *Bacteroides thetaiotaomicron*). The p-values are from Fisher's exact tests. These comparisons have been excerpted from the full set, which can be seen in Additional Table S2; q-values were calculated based on this full set of tests using the Benjamini-Hochberg method [73].

Background set	FDR	MODEL	p-value	odds ratio	q-value	significant
All tested (overall)	**5%**	**overall**	3.19×10^{-3}	**0.18**	7.65×10^{-3}	**TRUE**
Bacteroides (core or variable)	**5%**	**overall**	2.46×10^{-12}	**0.05**	2.46×10^{-11}	**TRUE**
Bacteroides (core only)	5%	overall	1.00	0.00	1.00	FALSE
Bacteroides (variable only)	**5%**	**overall**	7.20×10^{-4}	**0.13**	2.06×10^{-3}	**TRUE**
Bacteroides thetaiotaomicron only	**5%**	**overall**	2.96×10^{-6}	**0.09**	1.48×10^{-5}	**TRUE**
All tested (overall)	**25%**	**overall**	1.65×10^{-2}	**0.37**	3.41×10^{-2}	**TRUE**
Bacteroides (core or variable)	**25%**	**overall**	1.98×10^{-13}	**0.10**	2.37×10^{-12}	**TRUE**
Bacteroides (core only)	25%	overall	1.00	0.80	1.00	FALSE
Bacteroides (variable only)	**25%**	**overall**	4.12×10^{-4}	**0.18**	1.45×10^{-3}	**TRUE**
Bacteroides thetaiotaomicron only	**25%**	**overall**	6.04×10^{-7}	**0.18**	3.63×10^{-6}	**TRUE**
All tested (body site)	**5%**	**bodysite**	3.58×10^{-3}	**4.39**	8.27×10^{-3}	**TRUE**
Bacteroides (core or variable)	5%	bodysite	1.34×10^{-1}	2.00	2.44×10^{-1}	FALSE
Bacteroides (core only)	**5%**	**bodysite**	1.14×10^{-3}	**7.02**	2.98×10^{-3}	**TRUE**
Bacteroides (variable only)	5%	bodysite	6.25×10^{-1}	0.00	7.62×10^{-1}	FALSE
Bacteroides thetaiotaomicron only	5%	bodysite	2.77×10^{-1}	1.64	4.49×10^{-1}	FALSE
All tested (body site)	**25%**	**bodysite**	6.09×10^{-4}	**3.86**	1.86×10^{-3}	**TRUE**
Bacteroides (core or variable)	25%	bodysite	8.88×10^{-2}	1.78	1.72×10^{-1}	FALSE
Bacteroides (core only)	**25%**	**bodysite**	1.09×10^{-2}	**3.47**	2.33×10^{-2}	**TRUE**
Bacteroides (variable only)	25%	bodysite	4.51×10^{-1}	1.55	6.15×10^{-1}	FALSE
Bacteroides thetaiotaomicron only	25%	bodysite	4.38×10^{-1}	1.34	6.12×10^{-1}	FALSE

Supplementary figures/tables

Table S1: **Species prevalences, gut specificities, and Crohn's disease specificities for all genome clusters (species) tested.**

Table S2: **Full assessment of whether genes linked to microbial fitness in an *in vivo* experiment [43] were enriched for significant hits of the body site-specific and overall gut prevalence models.** The different sets of true positives were defined as: "Bacteroides" (genes in the screen significantly associated with fitness in all four strains), "BthetaDietIndep" (genes present in *Bacteroides thetaiotaomicron* that had diet-independent fitness effects in the screen), and "BthetaAny" (same, but for diet-dependent as well as -independent effects). The "background sets" were defined as follows: "all tested" (all gene families for which a phylogenetic model was fit), "Bacteroides (core or variable)" (all gene families with at least one representative in *Bacteroides* genome cluster pangenomes), "Bacteroides (core only)" (gene families that were present in all *Bacteroides* genome cluster pangenomes), "Bacteroides (variable only)" (gene families present in some but not all *Bacteroides* genomes clusters), and "Bacteroides thetaiotaomicron only" (only gene families present in *Bacteroides thetaiotaomicron*). Two false discovery rates for each model were tested (5% and 25%). Fisher tests yielded p-values that were then converted to q-values using the Benjamini-Hochberg approach [73].

Table S3: **SRA accession IDs used to estimate prevalence and environmental specificity scores.**

Figure S1: **Estimates of logit-gut prevalence (x-axis) vs. logit-gut specificity (y-axis), showing only modest correlation.**

Figure S2: **Laplacian regularization reduces noise in estimating $P(E|T)$.** Two species are compared, one that was infrequently observed in both Crohn's disease cases and controls (*Bacillus subtilis*, right) and one with a significant bias for Crohn's disease cases (*Bacteroides fragilis,* left). A) Total counts across subjects for *Bacillus subtilis* and *Bacteroides fragilis*. B) Likelihood function for $P(T|E)$, or prevalence in Crohn's disease. The maximum-likelihood value is given in the inset. C) Unregularized likelihood for $\text{logit}(P(E|T))$, or the environmental specificity of the microbe. Note that the maximum-likelihood value (inset) was actually almost twice as large for *Bacillus subtilis* as for *Bacteroides fragilis* despite the relative paucity of data for *B. subtilis* (compare Y-axes, which show that the distribution for *B. subtilis* is flatter). D) Laplace prior around $P(E) = 0.002$ with width parameter $b = 0.15$ (optimized using simulation). E) Log-likelihood plot for the posterior $P(E|T)$, obtained by taking the product of the prior distribution and the unregularized distribution. The maximum *a posteriori* estimates are the modes of these distributions (inset).

References

[1] Slack E, Hapfelmeier S, Stecher B, Velykoredko Y, Stoel M, Lawson MAE, et al. Innate and Adaptive Immunity Cooperate Flexibly to Maintain Host-Microbiota Mutualism. Science. 2009;325(5940):617–620. doi:10.1126/science.1172747.

[2] Atarashi K, Tanoue T, Shima T, Imaoka A, Kuwahara T, Momose Y, et al. Induction of Colonic Regulatory T Cells by Indigenous Clostridium Species. Science. 2011;331(6015):337–341. doi:10.1126/science.1198469.

[3] Mazmanian SK, Round JL, Kasper DL. A microbial symbiosis factor prevents intestinal inflammatory disease. Nature. 2008;453(7195):620–625. doi:10.1038/nature07008.

[4] Sassone-Corsi M, Raffatellu M. No vacancy: how beneficial microbes cooperate with immunity to provide colonization resistance to pathogens. Journal of immunology (Baltimore, Md : 1950). 2015;194(9):4081–7. doi:10.4049/jimmunol.1403169.

[5] Yano JM, Yu K, Donaldson GP, Shastri GG, Ann P, Ma L, et al. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. Cell. 2015;161(2):264–76. doi:10.1016/j.cell.2015.02.047.

[6] Peng L, Li ZR, Green RS, Holzman IR, Lin J. Butyrate Enhances the Intestinal Barrier by Facilitating Tight Junction Assembly via Activation of AMP-Activated Protein Kinase in Caco-2 Cell Monolayers. doiorg. 2009;139(9):1619–1625. doi:10.3945/jn.109.104638.

[7] Reber SO, Siebler PH, Donner NC, Morton JT, Smith DG, Kopelman JM, et al. Immunization with a heat-killed preparation of the environmental bacterium Mycobacterium vaccae promotes stress resilience in mice. Proceedings of the National Academy of Sciences. 2016;113(22):E3130–E3139. doi:10.1073/pnas.1600324113.

[8] Garrett WS, Gallini CA, Yatsunenko T, Michaud M, DuBois A, Delaney ML, et al. Enterobacteriaceae Act in Concert with the Gut Microbiota to Induce Spontaneous and Maternally Transmitted Colitis. Cell Host & Microbe. 2010;8(3):292–300. doi:10.1016/j.chom.2010.08.004.

[9] Kullberg MC, Ward JM, Gorelick PL, Caspar P, Hieny S, Cheever A, et al. Helicobacter hepaticus triggers colitis in specific-pathogen-free interleukin-10 (IL-10)-deficient mice through an IL-12- and gamma interferon-dependent mechanism. Infection and immunity. 1998;66(11):5157–66.

[10] Kostic A, Chun E, Robertson L, Glickman J, Gallini C, Michaud M, et al. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. Cell Host & Microbe. 2013;14(2):207–215. doi:10.1016/j.chom.2013.07.007.

[11] Bartlett JG. Clostridium difficile-associated Enteric Disease. Current infectious disease reports. 2002;4(6):477–483.

[12] van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, et al. Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. New England Journal of Medicine. 2013;368(5):407–415. doi:10.1056/NEJMoa1205037.

[13] Weingarden A, González A, Vázquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D, et al. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome. 2015;3(1):10. doi:10.1186/s40168-015-0070-0.

[14] Khanna S, Vazquez-Baeza Y, González A, Weiss S, Schmidt B, Muñiz-Pedrogo DA, et al. Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease. Microbiome. 2017;5(1):55. doi:10.1186/s40168-017-0269-3.

[15] Carvalho F, Koren O, Goodrich J, Johansson MV, Nalbantoglu I, Aitken J, et al. Transient Inability to Manage Proteobacteria Promotes Chronic Gut Inflammation in TLR5-Deficient Mice. Cell Host & Microbe. 2012;12(2):139–152. doi:10.1016/j.chom.2012.07.004.

[16] Chassaing B, Koren O, Carvalho FA, Ley RE, Gewirtz AT. AIEC pathobiont instigates chronic colitis in susceptible hosts by altering microbiota composition. Gut. 2014;63(7):1069–1080. doi:10.1136/gutjnl-2013-304909.

[17] Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nature Reviews Genetics. 2012;14(1):1–2. doi:10.1038/nrg3382.

[18] Kim PJ, Price ND. Genetic Co-Occurrence Network across Sequenced Microbes. PLoS Computational Biology. 2011;7(12):e1002340. doi:10.1371/journal.pcbi.1002340.

[19] Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. The ISME journal. 2017;11(1):248–262. doi:10.1038/ismej.2016.88.

[20] Zhao N, Chen J, Carroll I, Ringel-Kulka T, Epstein M, Zhou H, et al. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. The American Journal of Human Genetics. 2015;96(5):797–807. doi:10.1016/j.ajhg.2015.04.003.

[21] Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife. 2017;6. doi:10.7554/eLife.21887.

[22] Tang ZZ, Chen G, Alekseyenko AV, Li H. A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics. 2016;33(9):btw804. doi:10.1093/bioinformatics/btw804.

[23] Lozupone CA, Hamady M, Cantarel BL, Coutinho PM, Henrissat B, Gordon JI, et al. The convergence of carbohydrate active gene repertoires in human gut microbes. Proceedings of the National Academy of Sciences. 2008;105(39):15076–15081. doi:10.1073/pnas.0807339105.

[24] Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985;.

[25] Grafen A. The phylogenetic regression. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 1989;326(1233):119–57.

[26] Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. bioRxiv. 2017;.

[27] Ord TJ, Martins EP. Tracing the origins of signal diversity in anole lizards: phylogenetic approaches to inferring the evolution of complex behaviour. Animal Behaviour. 2006;71(6):1411–1429. doi:10.1016/j.anbehav.2005.12.003.

[28] Zaneveld JRR, Parfrey LW, Van Treuren W, Lozupone C, Clemente JC, Knights D, et al. Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation. Trends in microbiology. 2011;19(10):472–82. doi:10.1016/j.tim.2011.07.006.

[29] Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome research. 2016;26(11):1612–1625. doi:10.1101/gr.201863.115.

[30] Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008;103(482):681–686. doi:10.1198/016214508000000337.

[31] Gevers D, Kugathasan S, Denson L, Vázquez-Baeza Y, VanÂ Treuren W, Ren B, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe. 2014;15(3):382–392. doi:10.1016/j.chom.2014.02.005.

[32] Ives AR, Garland T. Phylogenetic Logistic Regression for Binary Dependent Variables. Systematic Biology. 2010;59(1):9–26. doi:10.1093/sysbio/syp074.

[33] Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. Nature. 2016;533(7604):543–546. doi:10.1038/nature17645.

[34] Swick MC, Koehler TM, Driks A. Surviving Between Hosts: Sporulation and Transmission. Microbiology spectrum. 2016;4(4). doi:10.1128/microbiolspec.VMBF-0029-2015.

[35] De Biase D, Pennacchietti E. Glutamate decarboxylase-dependent acid resistance in orally acquired bacteria: function, distribution and biomedical implications of the gadBC operon. Molecular Microbiology. 2012;86(4):770–786. doi:10.1111/mmi.12020.

[36] Srinivasa Rao PS, Lim TM, Leung KY. Functional genomics approach to the identification of virulence genes involved in Edwardsiella tarda pathogenesis. Infection and immunity. 2003;71(3):1343–51.

[37] Cotter PD, Gahan CG, Hill C. A glutamate decarboxylase system protects Listeria monocytogenes in gastric fluid. Molecular microbiology. 2001;40(2):465–75.

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

[39] Wargo MJ, Meadows JA. Carnitine in bacterial physiology and metabolism. Microbiology. 2015;161(6):1161–1174. doi:10.1099/mic.0.000080.

[40] Staley C, Weingarden AR, Khoruts A, Sadowsky MJ. Interaction of gut microbiota with bile acid metabolism and its influence on disease states. Applied microbiology and biotechnology. 2017;101(1):47–64. doi:10.1007/s00253-016-8006-6.

[41] Marques JC, Oh IK, Ly DC, Lamosa P, Ventura MR, Miller ST, et al. LsrF, a coenzyme A-dependent thiolase, catalyzes the terminal step in processing the quorum sensing signal autoinducer-2. Proceedings of the National Academy of Sciences. 2014;111(39):14235–14240. doi:10.1073/pnas.1408691111.

[42] Thompson J, Oliveira R, Djukovic A, Ubeda C, Xavier K. Manipulation of the Quorum Sensing Signal AI-2 Affects the Antibiotic-Treated Gut Microbiota. Cell Reports. 2015;10(11):1861–1871. doi:10.1016/j.celrep.2015.02.049.

[43] Wu M, McNulty NP, Rodionov DA, Khoroshkin MS, Griffin NW, Cheng J, et al. Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. Science. 2015;350(6256):aac5992–aac5992. doi:10.1126/science.aac5992.

[44] Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nature Reviews Microbiology. 2016;14(8):508–522. doi:10.1038/nrmicro.2016.83.

[45] Shin NR, Whon TW, Bae JW. Proteobacteria: microbial signature of dysbiosis in gut microbiota. Trends in biotechnology. 2015;33(9):496–503. doi:10.1016/j.tibtech.2015.06.011.

[46] Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. The New England journal of medicine. 2016;375(24):2369–2379. doi:10.1056/NEJMra1600266.

[47] Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology. 2014;32(8):822–828. doi:10.1038/nbt.2939.

[48] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology. 2014;32(8):834–841. doi:10.1038/nbt.2942.

[49] Glasser AL, Boudeau J, Barnich N, Perruchot MH, Colombel JF, Darfeuille-Michaud A. Adherent invasive Escherichia coli strains from patients with Crohn's disease survive and replicate within macrophages without inducing host cell death. Infection and immunity. 2001;69(9):5529–37.

[50] Barnich N, Boudeau J, Claret L, Darfeuille-Michaud A. Regulatory and functional co-operation of flagella and type 1 pili in adhesive and invasive abilities of AIEC strain LF82 isolated from a patient with Crohn's disease. Molecular Microbiology. 2003;48(3):781–794. doi:10.1046/j.1365-2958.2003.03468.x.

[51] Small CLN, Reid-Yu SA, McPhee JB, Coombes BK. Persistent infection with Crohn's disease-associated adherent-invasive Escherichia coli leads to chronic inflammation and intestinal fibrosis. Nature Communications. 2013;4:1957. doi:10.1038/ncomms2957.

[52] Lawley TD, Klimke WA, Gubbins MJ, Frost LS. F factor conjugation is a true type IV secretion system. FEMS microbiology letters. 2003;224(1):1–15.

[53] Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proceedings of the National Academy of Sciences. 2012;109(4):1269–1274. doi:10.1073/pnas.1113246109.

[54] Ives AR, Helmus MR. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs. 2011;81(3):511–525. doi:10.1890/10-1264.1.

[55] Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(45):19126–31. doi:10.1073/pnas.0906412106.

[56] Winter DJ. rentrez: An R package for the NCBI eUtils API. 2017;doi:10.7287/peerj.preprints.3179v2.

[57] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–2461. doi:10.1093/bioinformatics/btq461.

[58] Meyer F, Overbeek R, Rodriguez A. FIGfams: yet another set of protein families. Nucleic acids research. 2009;37(20):6643–54. doi:10.1093/nar/gkp698.

[59] Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research. 2014;42(Database issue):D581–91. doi:10.1093/nar/gkt1099.

[60] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010;5(3):e9490. doi:10.1371/journal.pone.0009490.

[61] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792–1797. doi:10.1093/nar/gkh340.

[62] Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–290. doi:10.1093/bioinformatics/btg412.

[63] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99–103. doi:10.1038/nature12198.

[64] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55–60. doi:10.1038/nature11450.

[65] Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics. 2013;14(1):19. doi:10.1186/1471-2105-14-19.

[66] Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Research. 2011;39(Database):D28–D31. doi:10.1093/nar/gkq967.

[67] Kappelman MD, Rifas-Shiman SL, Kleinman K, Ollendorf D, Bousvaros A, Grand RJ, et al. The Prevalence and Geographic Distribution of Crohn's Disease and Ulcerative Colitis in the United States. Clinical Gastroenterology and Hepatology. 2007;5(12):1424–1429. doi:10.1016/j.cgh.2007.07.012.

[68] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65. doi:10.1038/nature08821.

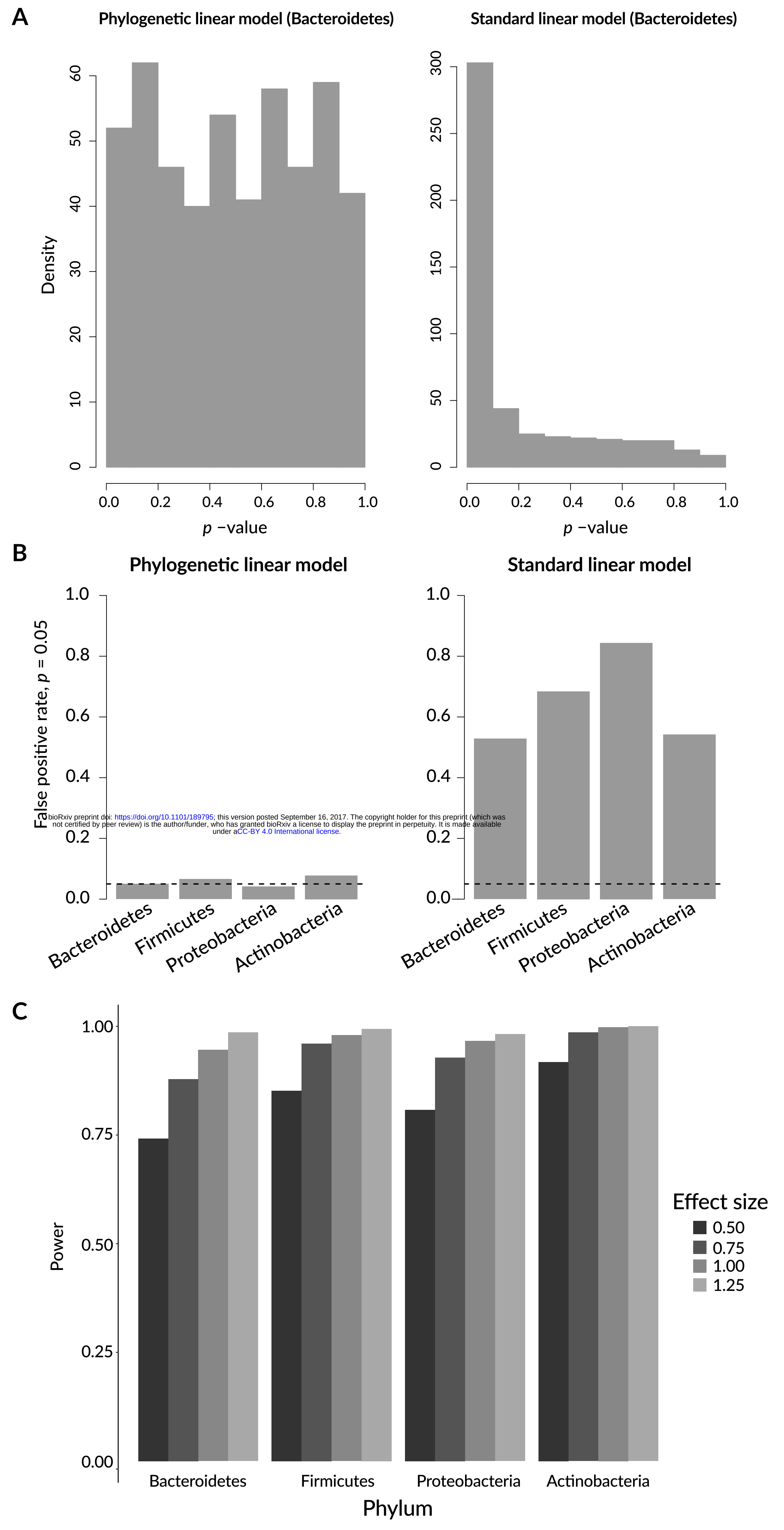
[69] si Tung Ho L, Ané C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. Systematic Biology. 2014;63(3):397–408. doi:10.1093/sysbio/syu005.

[70] Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control; 2015. Available from: `http://github.com/jdstorey/qvalue`.

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

[72] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Research. 2005;33(17):5691–5702. doi:10.1093/nar/gki866.

[73] Hochberg Y, Benjamini Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;1:289–300.



A

Glutamine synthetase type III, GlnN (EC 6.3.1.2)

Firmicutes

linear

Acidaminococcus intestinii

Dorea sp. 42_7

Pseudobutyrivibrio sp. MD2005

Lachnospiraceae bacterium 3−1

Butyrivibrio sp. AC2005

Ruminococcus flavefaciens

Clostridiales bacterium

Anaerococcus tetradius

Clostridium bartlettii

Anaerofustis stercorihominis

Erysipelotrichaceae bacterium 21_3

Erysipelotrichaceae bacterium NK3D112

B

Dihydroorotate dehydrogenase electron transfer subunit (EC 1.3.3.1)

Bacteroidetes

Prevotella amnii

Prevotella bivia

Prevotella sp. AGR2160

Prevotella copri

Prevotella sp. FD3004

Bacteroides ovatus

Bacteroides caccae

Bacteroides reticulotermitis

Bacteroidaceae bacterium MS4

Porphyromonas sp. 31_2

Tannerella forsythia

Butyricimonas virosa

Prolixibacter bellariivorans

Alistipes senegalensis

Alistipes indistinctus

C

Stage 0 sporulation two−component response regulator (Spo0A)

Firmicutes

phylogenetic

Firmicutes bacterium ZOR0006

Turicibacter sanguinis

Bacillus sp. NSP22.2

Salimicrobium sp. MJ3

Lachnospiraceae bacterium AC2028

Lachnospiraceae bacterium VE202−12

butyrate−producing bacterium SS3/4

Oribacterium sp. P6A1

Eubacterium desmolans

Ruminococcaceae bacterium AB4001

Pseudoramibacter alactolyticus

Anaerofustis stercorihominis

Erysipelotrichaceae bacterium 21_3

Erysipelotrichaceae bacterium

Holdemania filiformis

Sharpea azabuensis

Clostridium spiroforme

gut prevalence

2% 5% 12%

D

Glutamate decarboxylase (EC 4.1.1.15)

Bacteroidetes

Bacteroides fluxus

Bacteroides helcogenes

Bacteroides ovatus

Bacteroides vulgatus

Bacteroidetes bacterium oral taxon 272

Bacteroides paurosaccharolyticus

Porphyromonas sp. oral taxon 278

Porphyromonas cangingivalis

Parabacteroides johnsonii 9

Tannerella forsythia

Odoribacter splanchnicus

Prolixibacter bellariivorans

Alistipes indistinctus

Alistipes sp. ZOR0009

gene presence

absent present

E

Ives−Gardland α

Bacteroidetes

Firmicutes

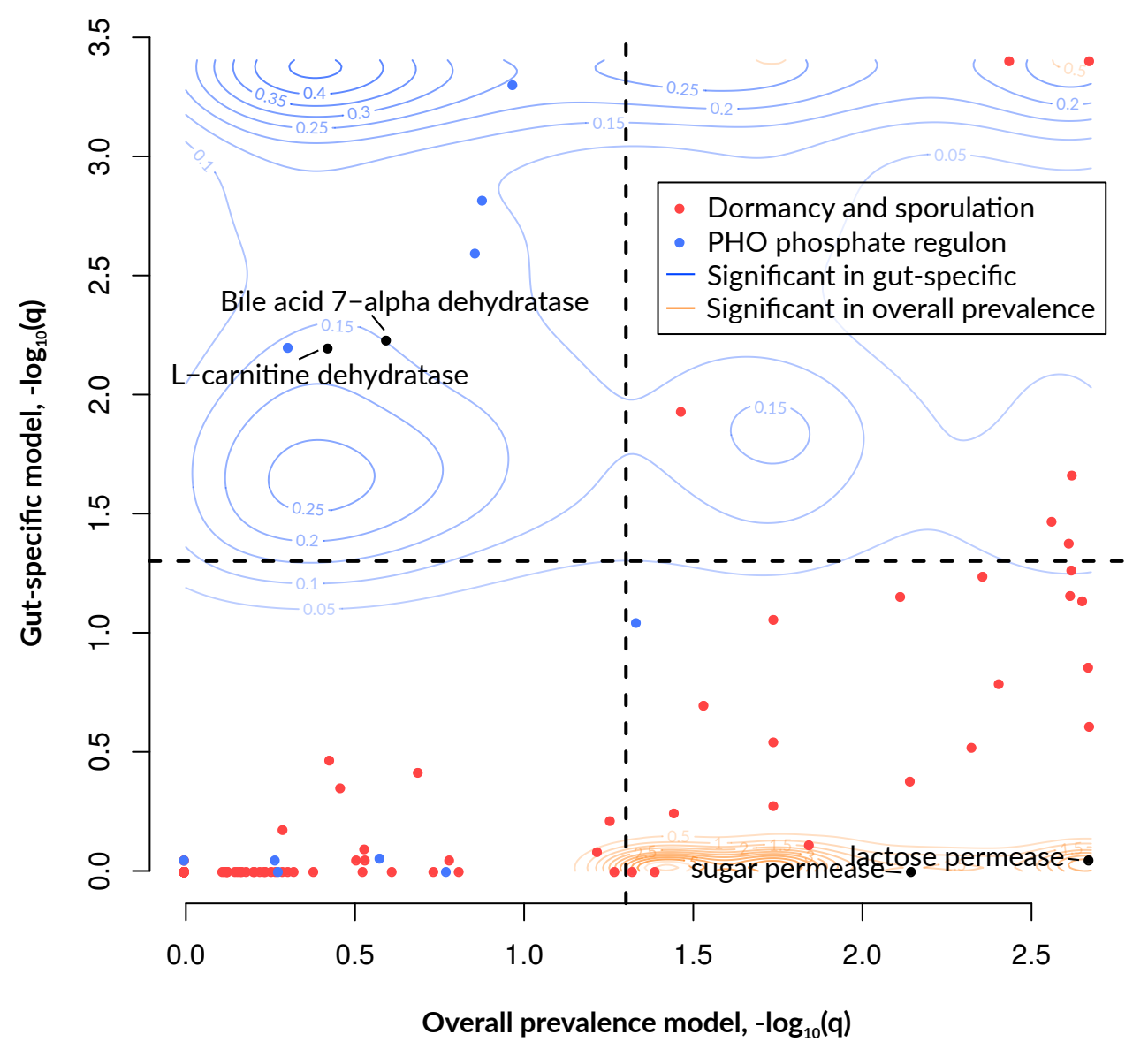
Proteobacteria

Actinobacteria

0 20 40

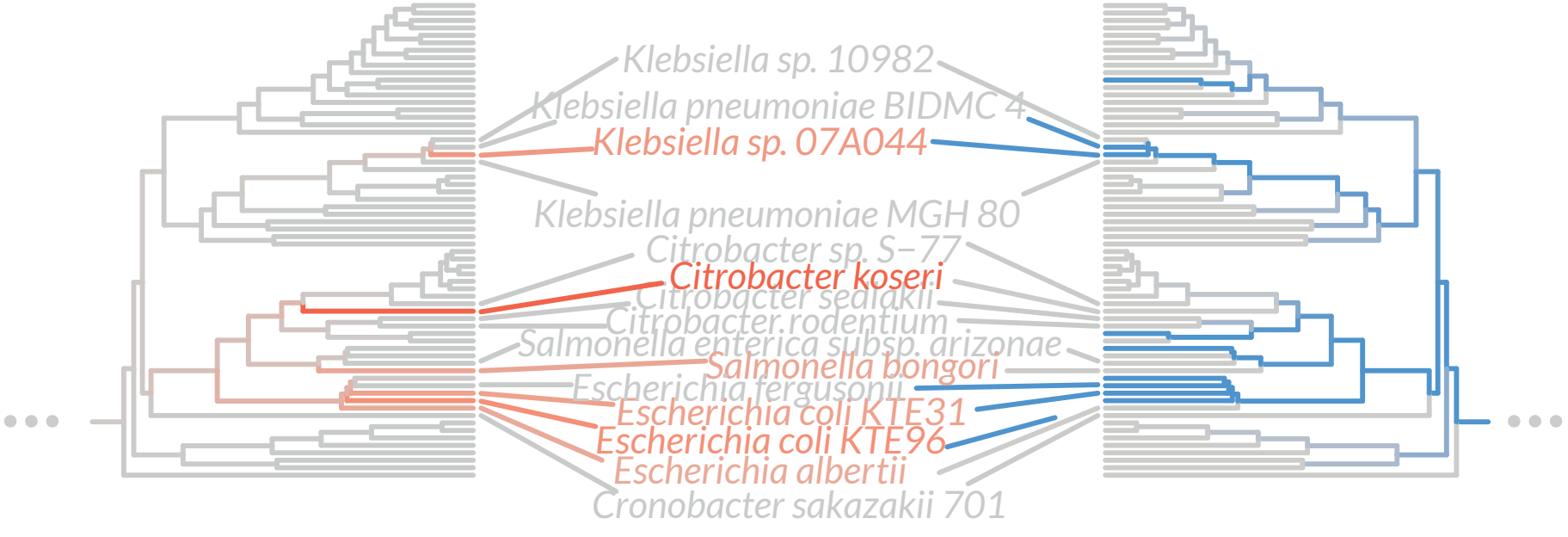
linear phylogenetic

method


Dormancy and sporulation
PHO phosphate regulon
Significant in gut-specific
Significant in overall prevalence
Bile acid 7−alpha dehydratase
L−carnitine dehydratase
sugar permease
lactose permease
Gut-specific model, -log10(q)
Overall prevalence model, -log10(q)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Conjugation protein TraR

Proteobacteria



P(CD|taxon)

0.002 **0.008**

gene presence

absent **present**