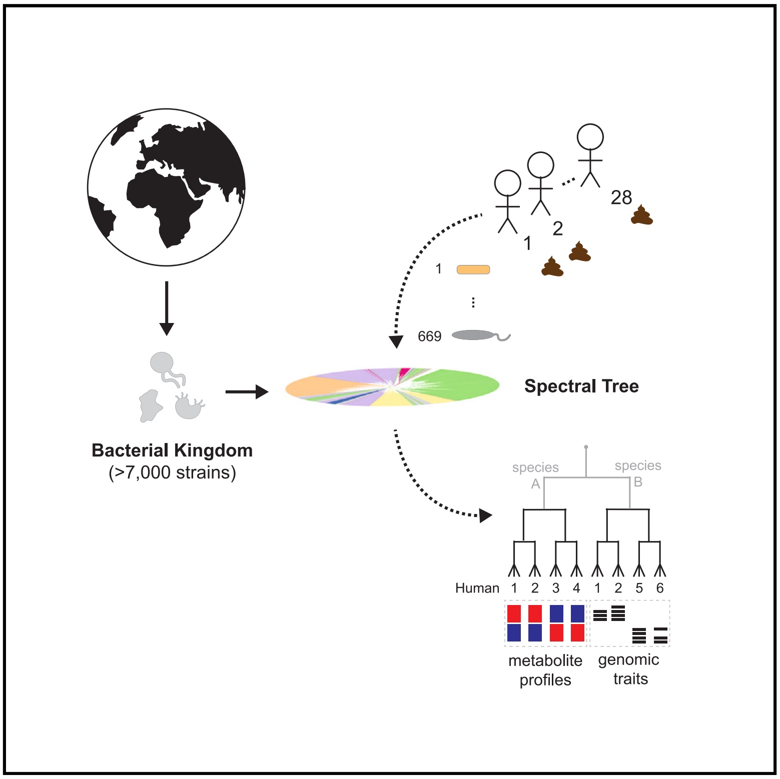
Cell Systems

Article

Subspecies phylogeny in the human gut revealed by co-evolutionary constraints across the bacterial kingdom

Graphical abstract



Highlights

- The spectrum of all principal components is a tree of nested constraint (Spectral Tree)
- The Spectral Tree of bacteria defines a latent space for evaluating strain banks
- Characterizing 669 human gut commensal strains revealed extensive subspecies phylogeny
- Models built from the Spectral Tree relate strain genotype with metabolic phenotype

Authors

Benjamin A. Doran, Robert Y. Chen, Hannah Giba, ..., Ashley Sidebottom, Eric G. Pamer, Arjun S. Raman

Correspondence

araman@bsd.uchicago.edu

In brief

Phylogenetic constraint can be inferred from patterns of co-evolution typically discarded as noise (Spectral Tree). The Spectral Tree of the kingdom Bacteria reveals the presence of extensive subspecies phylogeny among human gut strains and aids in relating genotype with phenotype. Our findings motivate defining strains according to their inferred co-evolutionary constraint.

Doran et al., 2025, Cell Systems *16*, 101167
February 19, 2025 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2024.12.008

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

Subspecies phylogeny in the human gut revealed by co-evolutionary constraints across the bacterial kingdom

Benjamin A. Doran,[1,2] Robert Y. Chen,[3] Hannah Giba,[1,4] Vivek Behera,[5] Bidisha Barat,[1] Anitha Sundararajan,[1] Huaiying Lin,[1] Ashley Sidebottom,[1] Eric G. Pamer,[1,5] and Arjun S. Raman[1,4,6,7,*]
[1]Duchossois Family Institute, University of Chicago, Chicago, IL 60637, USA
[2]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637, USA
[3]Department of Psychiatry, University of Washington, Seattle, WA 98195, USA
[4]Department of Pathology, University of Chicago, Chicago, IL 60637, USA
[5]Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[6]Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL 60637, USA
[7]Lead contact
*Correspondence: araman@bsd.uchicago.edu

https://doi.org/10.1016/j.cels.2024.12.008

SUMMARY

The human gut microbiome contains many bacterial strains of the same species (“strain-level variants”) that shape microbiome function. The tremendous scale and molecular resolution at which microbial communities are being interrogated motivates addressing how to describe strain-level variants. We introduce the “Spectral Tree”—an inferred tree of relatedness built from patterns of co-evolutionary constraint between greater than 7,000 diverse bacteria. Using the Spectral Tree to describe over 600 diverse gut commensal strains that we isolated, whole-genome sequenced, and metabolically profiled revealed (1) widespread phylogenetic structure among strain-level variants, (2) the origins of subspecies phylogeny as a shared history of phage infections across humans, and (3) the key role of inter-human strain variation in predicting strain-level metabolic qualities. Overall, our work demonstrates the existence and metabolic importance of structured phylogeny below the level of species for commensal gut bacteria, motivating a redefinition of individual strains according to their evolutionary context. A record of this paper’s transparent peer review process is included in the supplemental information.

INTRODUCTION

Microbial communities (“microbiomes”) are ubiquitous across diverse environments, spanning oceans to individual humans.[1–4] One microbiome relevant to human health is the gut microbiome: the trillions of microorganisms residing along the intestinal tract of humans.[5,6] A number of studies have demonstrated the significance of the gut microbiota—the bacteria within the microbiome—for influencing host physiology and predilection for developing several diseases.[7] This has led to many efforts describing the composition of human gut microbiotas and understanding how composition affects microbiome function. As interrogating microbiomes has become easier, the incredible taxonomic complexity of microbiotas and associated functional consequences have become more appreciated than ever before.[8,9]

Studies focused on cataloging microbiota composition have revealed the extensive presence of strain-level variants: strains that belong to the same species but are genetically different.[10–12] Moreover, several case studies have highlighted the direct impact of individual strains on gut microbiome function and host health. For instance, reconstitution of the infant gut microbiota using *Bifidobacterium longum* subspecies *infantis*—a strain of *B. longum* that metabolizes human milk oligosaccharides—has been shown to repair intestinal inflammation due to acute malnutrition in humans.[13–15] In another example, profiling different strains of *Bacteroides ovatus* showed differential capacity in inducing immunoglobulin A levels.[16] With respect to influence on bacterial fitness in the gut, interrogation of *Bacteroides* and *Parabacteroides* strains revealed strain-level preferences for binding polysaccharides.[17] Many similar vignettes, from understanding the etiology of food-borne outbreaks to characterizing the repair of the gut microbiota following antibiotic exposure, have highlighted the functional role of strain-level variants.[18,19]

As the functional importance of bacterial strains has become increasingly appreciated, an outstanding question is how should an individual strain be described? If considering complete genomes at amino acid resolution, almost every newly procured strain is a strain-level variant because it is likely to be different in some way from other strains of the same species. If instead the whole genome is compressed into a more practically manageable description, like the 16S rDNA sequence or sets of marker genes, strains become collapsed into their phylogenetic description obscuring potentially important adaptive changes that make strains of the same species functionally different from one another. With respect to classifying strains by biological function, recent studies that created banks of sequenced and phenotyped gut bacterial strains have shown a common trend: phenotypic differences between strains of the same species are difficult to understand.[20–22] As an example, it has been shown that metabolic capacities of bacteria follow coarse phylogeny but that variation among individual strains within a species is mostly unrelated to metabolic variability.[20] Collectively, these observations have led to the status quo strategy to functionally interrogate each and every new strain because structure amongst strain-level variants, i.e. "subspecies phylogeny," is difficult to ascertain and unlikely to be associated with strain-level phenotype.

A key limitation of performing comparative analysis on strain-level variants within strain banks is the tremendous degree of phylogenetic under-sampling compared with the bacterial tree of life. Strain banks usually contain strains from a specific econiche—only from the human gut for instance—and therefore reflect a small portion of phylogenetic diversity. While carefully curated reference trees of bacteria have been suggested as constructs to address phylogenetic limitations of strain banks, these trees are often also subset to strains from the specific econiche of interest. This limitation skews our understanding of gene content that is under selective pressure and likely associated with conserved phenotypes versus gene content that is allowed to vary and likely associated with the ability to adapt to different econiches.[23]

The existence of large databases of sequenced bacteria motivated a hypothesis that we tested here. Namely, that by leveraging the vast diversity of sequenced strains procured from many different environments in an unbiased way, we could better resolve evolutionary relationships. That is, constraints gleaned from the evolutionary record across the kingdom Bacteria could be used as a Bayesian prior for contextualizing differences between strains procured from a single environment—the human gut.[24,25] If achieved, this space of evolutionary relationships may (1) resolve fine-grained differences between strains of the same species, (2) allow testing whether phenotypic qualities specific to strain-level variants could be learned from genetic information, and (3) be a general construct for characterizing bacterial strains that could be dynamically updated as more strains are collected and sequenced.

We created a strain bank of 669 gut commensal strains that were isolated from fecal samples collected across 28 healthy human donors, whole-genome sequenced, and metabolically profiled. Consistent with previous studies, traditional analysis of this strain bank using 16S rDNA sequence, and well-known, standard sets of marker genes could not resolve genomic differences between strain-level variants or their associated metabolic qualities. We therefore developed an approach for inferring evolutionary distance between bacteria based on patterns of genomic covariation. Key to developing our approach was the theoretical finding that the whole spectrum of principal components (PCs) ("eigenspectrum") measured across extant diversity, including components typically discarded as noise, encoded a tree of relatedness that matched how two species co-evolve through sequential diversifications. Thus, covariation among extant diversity reflected constraints on co-evolution ("co-evolutionary constraint"). As our statistical approach was computationally fast, we applied it across >7,000 non-redundant bacterial proteomes isolated from many diverse environments to form a Spectral Tree of bacterial relatedness. We found that the Spectral Tree closely resembled known patterns of bacterial phylogenies. Examining our strain bank within the structure of the Spectral Tree revealed widespread subspecies phylogeny across gut commensal strains. Functional analysis showed that subspecies phylogeny was driven by a history of host phage exposure among groups of donors and was associated with a loss of well-conserved, biologically important genetic machinery. Finally, we used the Spectral Tree to predict strain metabolic capacity, finding that sampling strain-level variants among different donors (inter-donor) was key for building accurate predictive models of metabolism compared with strain-level variants procured from the same donor (intra-donor). We found this result was due to "inter-donor" strain-level genomic differences being substantially greater than "intra-donor" strain-level genomic differences.

Together, our findings demonstrate the existence of functionally significant subspecies bacterial phylogeny in the human gut revealed from analysis of co-evolution across the bacterial kingdom. Our work motivates a reparameterization from strain genomes to describing strains by their evolutionary context.

RESULTS

A bank of 669 metabolically profiled human gut commensal strains

We isolated and sequenced over 1,000 commensal bacterial strains from the feces of 28 healthy human volunteers. Our resulting bank of gut commensal strains ("commensal strain bank" from here on) comprised 669 diverse strains that we whole-genome sequenced (Figure S1A; Table S1) (STAR Methods). The commensal strain bank was enriched for gram-negative anaerobes within *Lachnospiraceae*, *Bacteroidaceae*, and *Bifidobacteriaceae* families (Figure S1B). We created phylogenetic trees of our strain bank defined by (1) the 16S rDNA sequence and (2) 120 proteins used to create the phylogenetic relationships in Genome Taxonomy Database (GTDB) ("Bac120")—the state-of-the-art database widely used for phylogenetic determination of bacterial strains (STAR Methods).[26] We found that both phylogenetic trees robustly defined coarse phylogenetic differences but were unable to resolve differences between strains belonging to the same species (Figure S1C).

We also metabolically profiled all strains within the strain bank across 50 targeted metabolites comprising amino acids, aromatics, branch-chained fatty acids, indoles, phenolic aromatics, and short-chain fatty acids (SCFAs) (Table S2) (STAR Methods). Metabolite concentrations relative to a standard in media (Brain Heart Infusion media supplemented with cysteine [BHIS]) without bacterial culture were measured. These metabolites were chosen to be profiled because they reflect particularly salient metabolites with respect to commensal bacterial fitness, human gut microbiome function, and interaction of the microbiome with the host.[27,28] Moreover, unlike other molecular signatures that are important but unable to be resolved at sufficiently high resolution for quantitative comparative studies like complex polysaccharides, each of these metabolites was associated with a unique mass-to-charge ratio, thereby facilitating comparative metabolomics. We found extensive variation among the metabolic capacity of strains from the same species (Figures S1D and S2). Together with the inability of canonical phylogenetic analysis to resolve strain-level variation, our findings were consistent with previously published studies illustrating the difficulty in relating strain-level variants with their metabolic capacity.[20]

Defining evolutionary distance using spectral inference

To test our hypothesis that evolutionary relationships across a wealth of sequenced bacteria could aid in revealing strain-level genomic differences within our strain bank, we turned to a large database of sequenced non-redundant bacterial strains procured across a diversity of environments. Previous work from our laboratory described a phylogenomic analysis of the kingdom Bacteria using all reference proteomes in the UniProt database (>7,000 strains in total).[29,30] Analysis of this database illustrated that co-evolutionary patterns of proteome variation defined a hierarchy of phylogeny. Major PCs clustered bacteria belonging to the same phylum, deeper components class, and so on until species. Building upon this finding, we reasoned that the whole PC spectrum of bacterial co-evolution defined across the UniProt database, including PCs typically discarded as noise, may be useful for inferring evolutionary distances between bacteria and resolve fine-grained differences between strains of the same species. This idea motivated creating a metric of evolutionary distance between extant taxa based on statistical patterns of proteome co-evolution.

Developing a definition of evolutionary distance inferred from patterns of bacterial co-evolution first required studying simple, manageable models of the evolutionary process. Thus, we used "toy model" simulations of diversification and selection to explore how PCs could be used to define an evolutionary distance metric between taxa that are represented by a complex set of features, like a genome. An example of one such model is shown in Figure 1A. Here, a "parent" (root) was defined by a set of features ("genotype") and was subject to sequential, layered diversifications to create an alignment of taxonomic diversity. Traditionally, evolutionary simulations are performed using nucleotide identities (adenine ["A"], thymine ["T"], cytosine ["C"], guanine ["G"]) as the basis of variation. In our model, we defined the basis of variation to be "A"/"T" for simplicity, and we also encoded "A" as a "0" and "T" as a "1" basis so that quantitative analysis could be performed on the model. The resultant alignment was then subject to principal-component analysis (PCA) (Figure 1B). The output of PCA is a spectrum of "PCs" ordered by statistical scale: PC1 harbors the most amount of data variance, PC2 the second-most, and so on. Each taxon of the alignment and each feature describing a taxon contribute to each PC to a certain extent.

In general, the shallowest PCs (PC1, PC2, and PC3, for instance) are thought to contain relevant signal while deeper PCs are typically discarded as statistical noise. Our analysis of the alignment in Figure 1A illustrated a different finding. The taxa arising from a common broad layer of diversification contributed similarly to the first two PCs, while those arising from common finer layers of diversification (the second and third diversifications) contributed similarly to shallow as well as deeper PCs (Figure 1C). We therefore defined a metric termed "spectral distance" between two taxa:

$$SD_{ij}^{k} = \left| P_i^k - P_j^k \right| \qquad \text{(Equation 1)}$$

where $\boldsymbol{P_i^k}$ is the contribution of taxon $\boldsymbol{i}$ onto PC $\boldsymbol{k}$ and $\boldsymbol{P_j^k}$ is the contribution of taxon $\boldsymbol{j}$ onto PC $\boldsymbol{k}$. Plotting the cumulative spectral distance across all PCs, including those harboring a minutia of data variance (~1%), defined a tree-like hierarchical pattern of statistical similarity between taxa (Figure 1D). Moreover, we found that sequential PCs collectively described different layers of precedent diversifications. For instance, for the example shown in Figure 1A, PCs 5–8 harbored the same amount of data variance and collectively described the third layer of diversification from the parent, as observed in taxa "a" and "b" being maximally differentiated along PC7 (Figures 1C and 1D). Similarly, PCs 3 to 4 harbored the same amount of data variance and described the second layer of diversification from the parent, as observed in taxa "a" and "b" being maximally differentiated from taxa "c" and "d" along PC4 (Figures 1C and 1D). PCs 1 and 2 independently described the root and the first layer of diversification, respectively. The result from Figure 1D motivated grouping PCs into "spectral groups" that harbor the same percent data variance and then computing spectral distance across the spectral groups (Figure 1E, left) (STAR Methods). Hierarchical clustering of the resulting pairwise spectral distances across all taxa yielded a tree of taxonomic relatedness that we termed a Spectral Tree (Figure 1E, right). We found that for the case in Figure 1A, the Spectral Tree matched existing approaches of phylogenetic inference spanning maximum-likelihood and Bayesian methods (Figure 1F). The exception to this finding was Mr. Bayes, which resulted in a star-like pattern of inferred relatedness likely due to the limited size of the alignment.

We next tested whether Spectral Trees resulted in accurate taxonomic relationships across (1) different sizes of alignments and (2) different numbers of genotypic features used to describe each system in the alignment. Synthetic phylogenetic histories were created *in silico*, Spectral Trees were generated from the alignment of taxa for each synthetic dataset, and a measure of how accurately the resulting Spectral Tree captured the phylogenetic history in the dataset was then computed (STAR Methods). We found that for much of the parameter space we analyzed, the Spectral Trees closely resembled the ground-truth pattern of sequential diversifications (Figure S3). The exception to this trend was in the limit where the number of features was less than the number of taxa, in which case the Spectral Tree did not match the ground-truth tree. This is because in the regime where taxa outnumber the features used to describe taxa, the number of features describing each taxon is limited compared with the diversity of taxa available for sampling. As such, the information content of the set of features is "overwritten" by the diversity of taxa, thereby erasing patterns of covariation originating from phylogenetic histories—a scenario that the Spectral Tree is not designed to

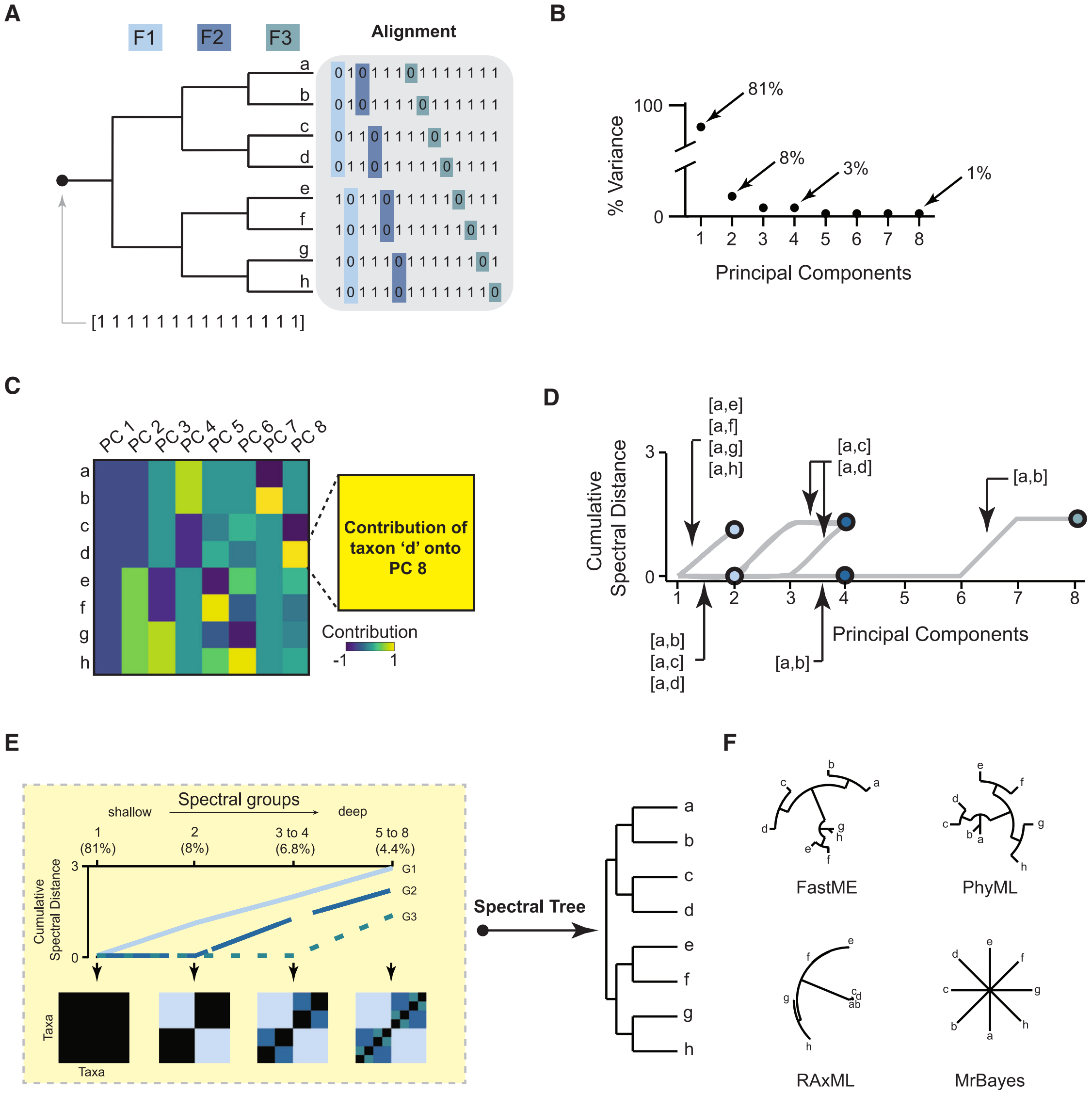


Figure 1. Defining a Spectral Tree from extant diversity: An example using a toy model

(A) An *in silico* model of sequential diversification. An ancestral "root" is defined by a 14-bit string of "1." Diversification through three generations creates an alignment of eight "taxa." Colored bits in the alignment match the color of the generation "F1," "F2," or "F3" at which variation from a "1" to a "0" was introduced.

(B) PCA of the alignment in (A) yields eight PCs. Percent variance harbored by each PC is shown.

(C) Contributions of each taxon in the alignment from (A) to each PC.

(D) Cumulative spectral distance (y axis) for all pairs of taxa that include taxon "a." The pattern of cumulative spectral distances resembles a tree-like distribution.

(E) PCs are grouped together based on their percent variance into "spectral groups." For each spectral group, spectral distances are computed between all pairs of taxa. Spectral distances between all pairs of taxa for each spectral group are displayed and black to blue pixel colors indicate low to high spectral distances. This information is used to create a rooted Spectral Tree.

(F) Unrooted trees resulting from phylogenetic inference methods applied to alignment in (A).

capture. A biological process consistent with this regime is when the recombination rate is extremely high relative to speciation events—a scenario that has been put forth as a plausible explanation for bacterial phylogenomic trends.[31]

Spectral Trees resolve convergent paths of diversification

Our results motivated characterizing situations in which Spectral Trees were distinct from current methods of phylogenetic inference. Analysis of Spectral Trees across a diversity of alignments illustrated that Spectral Trees were qualitatively distinct compared with existing methods of phylogenetic inference in cases of convergent processes. Convergent evolution involves two or more taxa possessing the same set of genomic traits through independent ancestral histories. These convergent histories vastly complicate phylogenetic inference because genomic diversity no longer increases in a predictable manner over evolutionary time. Disentangling evolutionary convergence

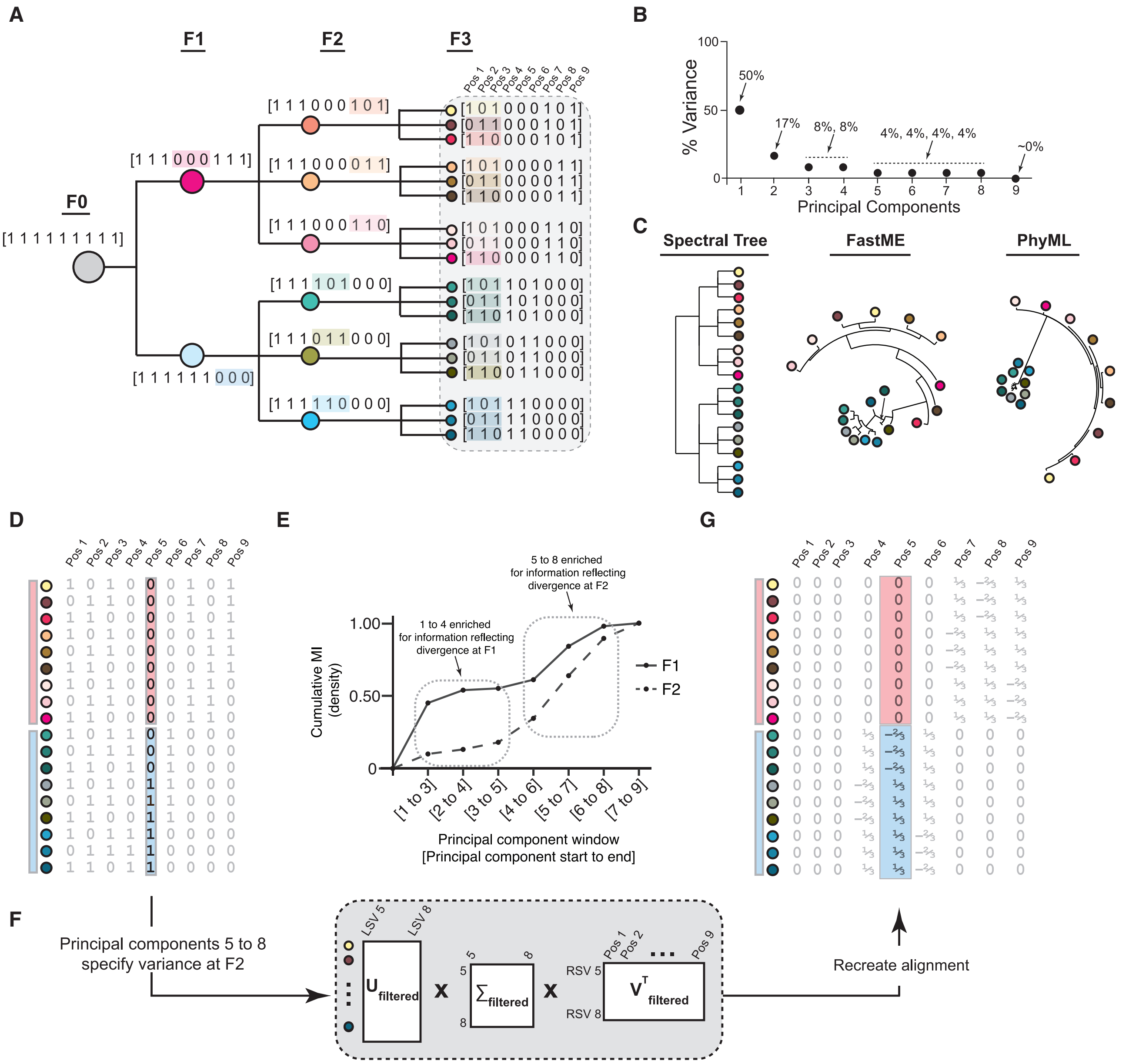


Figure 2. Spectral Trees resolve patterns of convergent diversification

(A) The root (F0) is defined by a 9-feature genotype of "1" and subject to three sequential diversification events (F1, F2, and F3), resulting in 18 taxa, and features can be either "1" or "0." Colored features in each generation correspond to variation from the previous generation.

(B) Scree plot of nine PCs describing alignment in (A).

(C) Rooted Spectral Tree, unrooted trees resulting from FastME and PhyML.

(D) Alignment of taxa with position 5 highlighted, and taxa are labeled by the group they belong to within the F1 generation (red or blue bar).

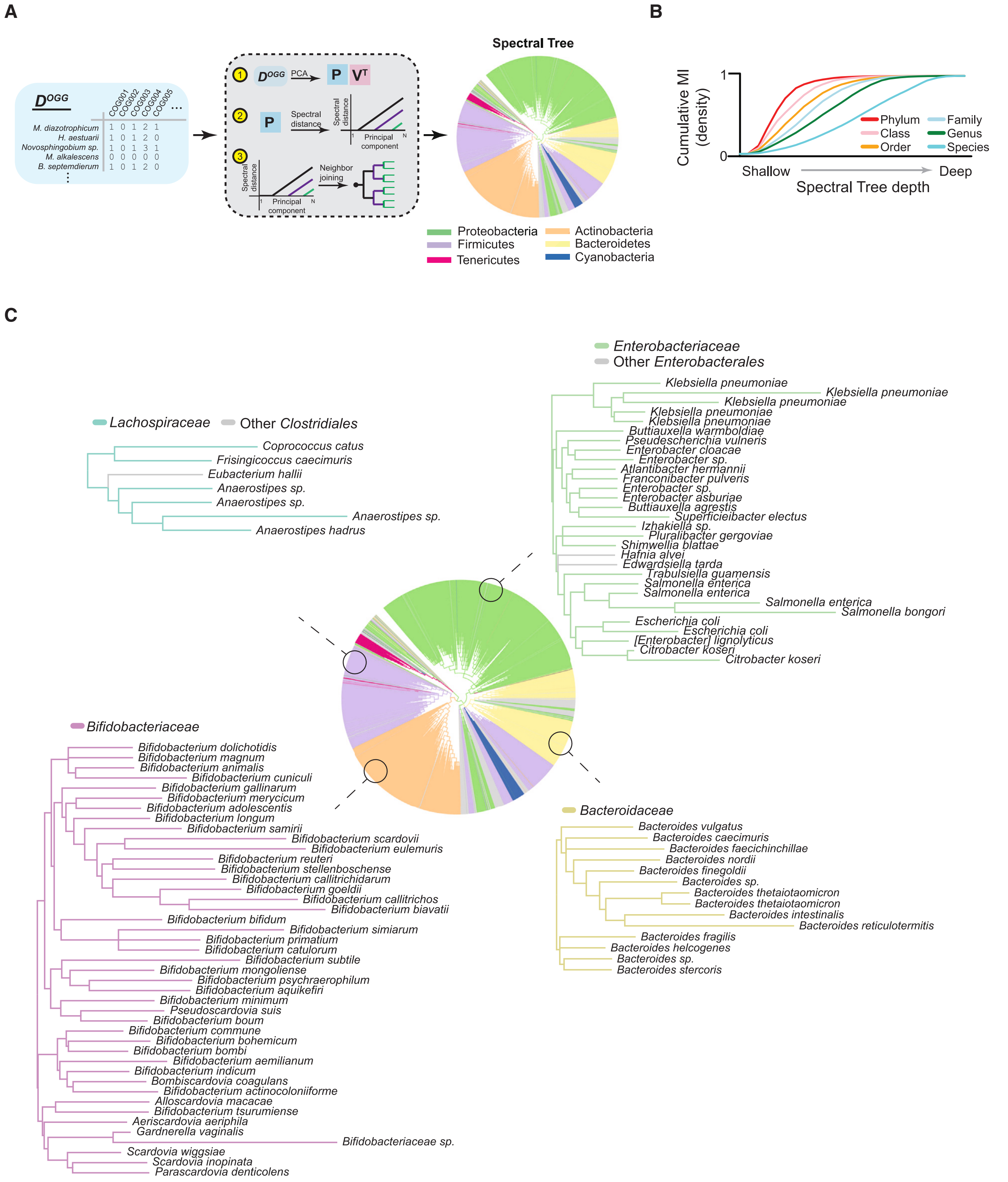
(E) Shared information (cumulative mutual information [MI], y axis) between clustering of taxa across sets of PCs (x axis) and clustering defined at the F1 or F2 generations (legend).

(F) Isolation of information contained in PCs 5 through 8 using singular value decomposition (SVD).

(G) Recreated alignment considering information contained only in PCs 5 through 8 with position 5 highlighted. Taxa are labeled in the same manner as (D).

statistically is particularly problematic when using single features (i.e., "gene markers") to model ancestral distance because such approaches, like maximum-likelihood or Bayesian methods, do not explicitly consider pairwise or higher-order epistasis between genes and the resulting complex contextual dependence of gene presence or absence on other genes.[32]

To understand why Spectral Trees were effective at dealing with evolutionary convergence, we interrogated a paradigmatic example of a representative trajectory (Figure 2A). In this example, the ancestral root ("F0") was followed by three sequential sets of diversifications leading to 18 diverse taxa. We found that the Spectral Tree of taxa correctly captured the



	COG001	COG002	COG003	COG004	COG005	...
M. diazotrophicum	1	0	1	2	1	
H. aestuarii	1	0	1	2	0	
Novosphingobium sp.	1	0	1	3	1	
M. alkalescens	0	0	0	0	0	
B. septemdierum	1	0	1	2	0	
⋮						

Figure 3. A Spectral Tree of 7,047 bacteria from UniProt

(A) Workflow for computing a Spectral Tree across all non-redundant bacterial strains within UniProt (n = 7,047). $\boldsymbol{D}^{\boldsymbol{OGG}}$ is the matrix of 7,047 UniProt reference bacterial strains (rows) annotated by their OGG content (columns) (blue box), and each entry is the number of sequences corresponding to a specific OGG within a

(legend continued on next page)

generative set of diversifications spanning the F1 and F2 generations, while application of other methods (FastMe and PhyML) did not (Figures 2B and 2C).

To better understand this result, we analyzed how features in the alignment were contributing to each PC. As an example, we considered position 5 in the alignment (Figure 2D). Position 5 was a "0" for all taxa arising from the top branch of F1 and a "0" for one-third of taxa arising from the bottom branch of F1. Thus, two separate contexts evolved independently, resulted in a "0" at position 5—an example of convergence. Using position 5 as a marker would therefore lead to an incorrect grouping of taxa arising from different histories together. This is a commonly encountered problem when considering each genomic feature in isolation of its genetic context. Generalizing this concept, for all positions in the alignment shown in Figure 2A, no single position was sufficient to describe a single diversification event.

Given this finding, we sought to elucidate where information regarding the different generations of diversifications lay across the set of PCs (STAR Methods). We found that the shallowest PCs were enriched for information regarding a shared history at the F1 generation, while the deepest set of PCs were enriched for information regarding a shared history at the F2 generation (Figure 2E). We then recreated the alignment using only information contained within the deepest PCs (Figure 2F) (STAR Methods). Focusing on position 5 again, we found that the value of position 5 was adjusted in the recreated alignment reflecting the separate, nested contexts of diversification (Figure 2G).

Historically, eigendecomposition—the spectral factorization technique underlying PCA—has been used as a form of dimension reduction: analyze only the shallowest PCs for significant biological trends. This use of PCA has stemmed from application of random-matrix theory (RMT) to biological data.[33] Our findings provide a contrasted result, demonstrating that the whole eigenspectrum (spectrum of PCs resulting from PCA) can encode a Spectral Tree and that the tree is the more complete dimension-reduced object. Through a detailed mathematical analysis, we found that the formation of hierarchy as represented in a Spectral Tree is guaranteed from performing eigendecomposition on related populations (supplemental experimental procedures Section 1) (Figure S4). Specifically, the deep PCs contain information that is nested within shallower PCs. Therefore, the qualities we found regarding convergent processes are not specific to the toy model in Figure 2A but rather are general properties of using the eigenspectrum to create trees of relatedness. A more detailed explanation of comparing standard methods of phylogenetic inference with creating Spectral Trees using eigendecomposition can be found in supplemental experimental procedures Section 2.

The Spectral Tree built from 7,047 bacterial strains in UniProt reflects known phylogenetic patterns

We sought to create a Spectral Tree for a large diversity of non-redundant bacterial strains representative of the kingdom Bacteria. We turned to the UniProt non-redundant database comprised of 7,047 strains for this task. To represent bacterial diversity in a more unbiased and complete manner compared with 16S or the set of Bac120 gene markers used to define GTDB, we annotated each bacterium by its orthologous gene group (OGG) content. OGGs are groups of proteins defined by the conservation pattern of their amino acid sequences and have been used previously for phylogenomic comparisons in bacteria.[30,34,35] Our strategy resembles that of pan-genomic analysis in analyzing the abundance of information within both "core" and "accessory" genomic regions.

We first tested whether building a Spectral Tree across thousands of reference proteomes in the UniProt database would be computationally feasible. We selected members of the class *Bacteroidia* ($n = 211$), order *Oceanospirillales* ($n = 103$), family *Rhodospirillaceae* ($n = 50$), and genus *Ruminococcus* ($n = 25$) annotated with 10,177 OGGs and found that computing a Spectral Tree required substantially less computational resources than existing methods of phylogenetic inference (Figure S5). This result motivated computing a Spectral Tree across thousands of non-redundant taxa—a goal that is not practically feasible with current approaches.

We constructed a Spectral Tree for the set of non-redundant bacterial proteomes in UniProt using an alignment of 7,047 bacteria annotated by their OGG content (Figure 3A; Table S3) (STAR Methods). Analyzing the Spectral Tree at the level of phylum showed that generally, groups of bacteria belonging to the same phylum clustered together. Phyla that were consistently monophyletic across GTDB and NCBI, such as Actinobacteria and Cyanobacteria, remained monophyletic in the Spectral Tree (Figures S6A and S6B).[26,36] Additionally, phylogenetic relationships between phyla were maintained. For instance, the Tenericutes were placed between Proteobacteria and Firmicutes—a phylogenetic relationship that has been previously described and represented in bacterial phylogenetic trees.[26,37] In another example, GTDB reclassified Proteobacteria from NCBI into Pseudomonodota and Desulfobacteria.[26] The Spectral Tree captured this reclassification (Figure S6C). However, there were notable instances where clusters of bacteria deviated from their known phylum-level designation. First, there were two groups of Firmicutes that were separated from each other and from the main group of Firmicutes. These two outgroups are the order *Bacillales*. This split is partially supported by GTDB's reassignment of the class Bacilli to phylum *Bacilliota*. The exception in our Spectral Tree is that the order *Lactobacillales* is considered to be more related to the main group of Firmicutes. Second, within the phylum Bacteroidetes, we observed two major separated classes—*Flavobacteriia* and *Cytophagia*. Third, we found the placement of bacteria belonging to phyla with representation of less than 100 total members was enriched in proximity to Bacteroidetes relative to other major phyla like Actinobacteria, Firmicutes, and Proteobacteria. These observations and discrepancies with respect to NCBI classification are

specific bacterial proteome. Gray box outlines the three computational steps for creating a Spectral Tree from $\boldsymbol{D}^{OGG}$. Spectral Tree is shown with leaves (each of the 7,047 bacteria) colored by phylum per NCBI.

(B) Information shared between clusters of the Spectral Tree, ordered from shallow to deep (x axis), and phylogenetic classification. Shallowest Spectral Tree clusters are enriched for grouping bacteria together by phyla, deepest clusters by species.

(C) Zoom-ins of Spectral Tree at specific bacterial families.

likely due to two factors: (1) the Spectral Tree was built from OGG frequency across the entire proteome rather than considering conserved marker genes and (2) our statistical framework incorporates and leverages epistasis between OGGs.

We next performed a systematic analysis of the phylogenetic distribution of bacteria across the Spectral Tree. To do this, we measured the mutual information between shared tree depth in the Spectral Tree and shared phylogeny (STAR Methods). This analysis revealed how much information about phylogenetic relationships was captured by respecting the clustering of bacteria created by the Spectral Tree. Our results showed that the topology of the tree matched a hierarchy of phylogeny: shallow to deep Spectral Tree clusters progressively grouped bacteria from the same phylum, class, order, family, genus, and species (Figure 3B). Moreover, zooming-in on specific bacteria belonging to families that are common in the human gut illustrated that the Spectral Tree captured known phylogenetic relationships at the species level (Figure 3C). Thus, while we observed certain exceptions to canonical phylogenetic classifications, overall the Spectral Tree recapitulated known evolutionary relationships across thousands of bacterial strains.

Using the Spectral Tree to resolve subspecies phylogeny within our strain bank

The Spectral Tree of relationships built using bacterial strains comprising the UniProt database was a statistical space that captured known evolutionary relationships. In this sense, we conceptualized the Spectral Tree as an evolutionarily relevant "latent space"—an abstract space where distance between objects scales with a desired property. In our case, the objects are bacterial proteomes, and the desired property is evolutionary relatedness. Using this conceptualization, new strains could be projected into the latent space thereby making the Spectral Tree a dynamic object capable of incorporating more strains to reflect the increasing corpus of bacterial sequencing data. We therefore saw this as an opportunity to characterize our gut commensal strain bank using the Spectral Tree.

We annotated all strains in our strain bank by their OGG content and projected each strain into the Spectral Tree (Figures 4A, S7, and S8; Table S4) (STAR Methods). We compared the distances of all pairs of strains in our strain bank that share the same genus or species designations computed from (1) the Spectral Tree, (2) the phylogenetic tree created from the 16S rDNA sequence, or (3) the phylogenetic tree created from Bac120. We found that for pairs of strains from the same species, the Spectral Tree uniquely resolved differences between our strains: the average relative distance of strain pairs based on 16S and Bac120 trees was zero, while the same distribution based on the Spectral Tree was bimodal (Figure 4B). We also found that creating a Spectral Tree of the commensal strain bank without considering the UniProt database yielded significantly less separation of strains at phylogenetic scales that were coarser than species-level designations (Figure S9). Collectively, these results suggested that using the Spectral Tree built from the UniProt database as a latent space for characterizing bacteria-resolved phylogenetic relationships, from broad to subspecies-level phylogenetic differences, between strains in our commensal strain bank.

We next sought to interrogate the structure of strain-level variation within the Spectral Tree. Focusing on the group of 41 *Mediterraneibacter gnavus* strains in our strain bank, we found that the Spectral Tree defined phylogenetic structure through species-level designation but also showed statistically significant non-random clustering among strain-level variants. Notably, we found a direct relationship between the structure of strain-level variation and donors from which strains were collected (Figure 4C, upper). In another example, the 27 strains of *Bacteroides uniformis* illustrated the same trend of being clustered by donor origin (Figure 4C, lower). This result suggested that the Spectral Tree was defining subspecies phylogenetic structure based on proteome differences in strains associated with individual donors.

To test the generality of this result across the entire commensal strain bank, we computed the mutual information between strain clusters defined across the Spectral Tree and whether the clusters shared the same phylogenetic designation or donor origin. We found that the pattern of strain clustering across the tree reflected a distinct biological order: shallow clusters reflected broad phylogenetic differences, deeper clusters reflected finer phylogenetic differences, and the deepest clusters reflected variation between strains of the same species but isolated and cultured from different donors (Figure 4D).

Thus, our results illustrated two related findings. First, the Spectral Tree revealed a phylogenetic structure present below the level of species. Second, this subspecies phylogenetic structure was associated with diversification in the econiche of different humans.

In totality, the Spectral Tree contained 41 layers. The layer at which subspecies phylogeny was defined was layer 26 (Figure 4D). As the Spectral Tree was built from >7,000 PCs spanning over 10,000 OGGs, we sought to understand how the Spectral Tree organizes the vast genomic information used as input. To delineate the pattern of OGGs that define hierarchical relationships in the Spectral Tree, we identified OGGs that were significantly differentially abundant between daughter branches of a given cluster (Figure S10). Interrogating the pattern of OGGs across clusters in the Spectral Tree, we found that the Spectral Tree is organized through nested genomic variation. For instance, variation in OGGs defining the second layer of the Spectral Tree was nested within OGGs whose variation defined a cluster in the first layer. This hierarchical pattern continued until the last layer of the tree (Figure S11A). Crucially, this property of nestedness enabled explicitly identifying genomic differences that distinguished clusters of strains—a property we used to functionally characterize subspecies phylogeny as described next (Figure S11B).

Functional and evolutionary characterization of subspecies phylogeny

What are the origins of structured phylogeny below the level of species? We used the Spectral Tree to better understand drivers of subspecies phylogeny within our strain bank. As an example, our strain bank contained 20 strains of *Eubacterium rectale* (also called *Agathobacter rectalis*) collected from several donors. We isolated the Spectral Tree branch that separated different groups of *E. rectale* strains. As expected per our results in Figure 4D, the groups of strains clustered by donors from which they were isolated (MSK17 and MSK22 versus MSK16, MSK13, and MSK9; "MSK" stands for Memorial Sloan Kettering, one of the hospitals

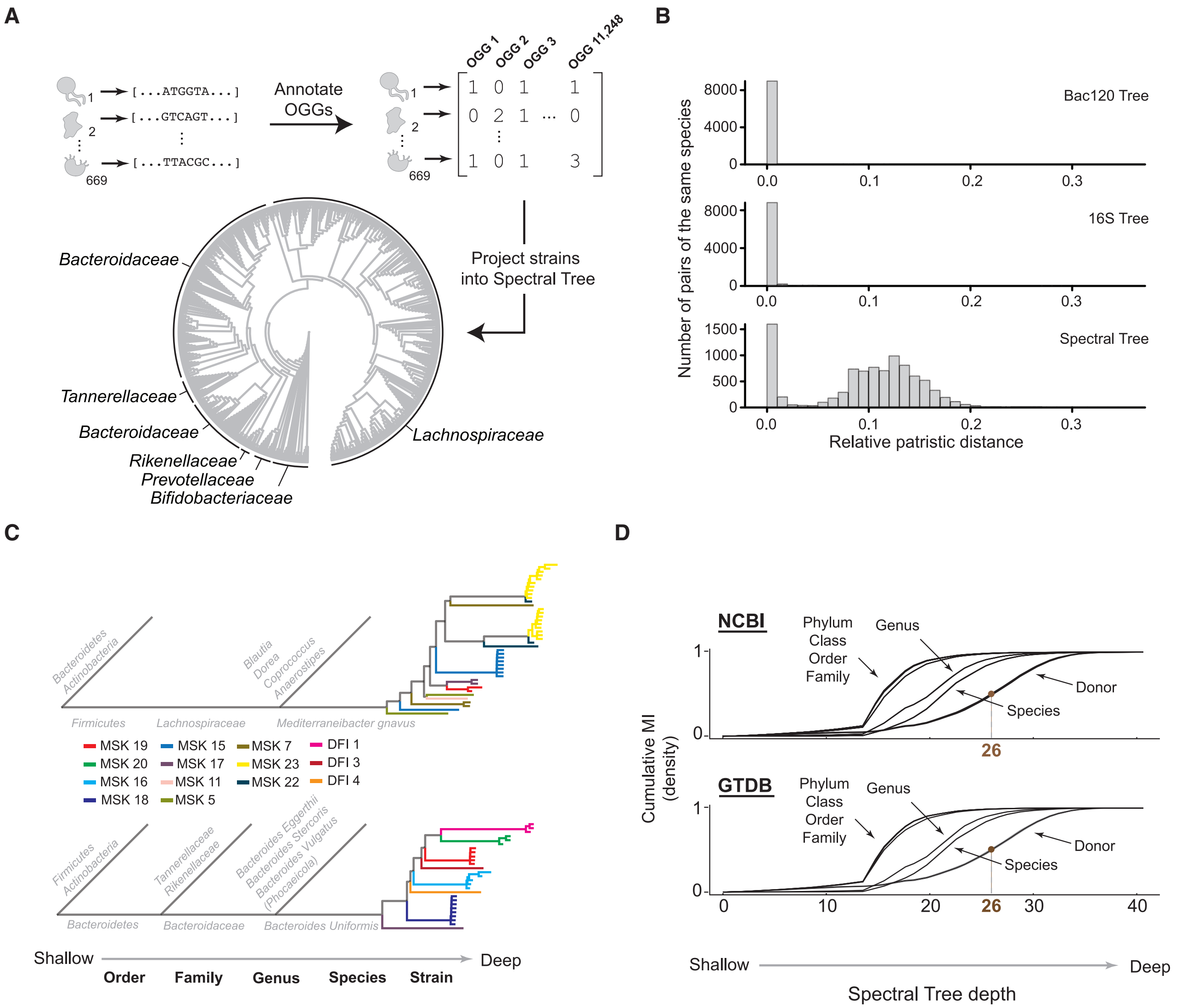


Figure 4. The Spectral Tree reveals subspecies phylogeny in commensal strain bank

(A) Workflow for projecting commensal strain bank into the Spectral Tree (see Figure S7 for detailed steps).

(B) Distributions of relative distances for all strain pairs in the commensal strain bank that are of the same species. Relative distance is defined by either (1) the Bac120 phylogenetic tree (top), (2) the 16S phylogenetic tree (middle), or (3) the Spectral Tree (bottom).

(C) Following strains of *M. gnavus* (upper) and *B. uniformis* (lower) from shallow to deep branches of the Spectral Tree. Each leaf is a strain colored by the identity of the human donor from which the strain was collected (see color key, MSK indicates donor from Memorial Sloan Kettering Hospital, and DFI indicates donor from Duchossois Family Institute, University of Chicago).

(D) Information shared between phylogenetic designation (NCBI or GTDB database) or donor origin and depth of strain cluster in Spectral Tree (x axis). Tree depth at which 50% of cumulative information regarding shared donor identity is delineated (brown).

from which donors were recruited and fecal samples were isolated) (Figure 5A). Differences in OGGs between the strains of *E. rectale* illustrated a pattern of mutually exclusive presence or absence. Strains isolated from donors MSK22 and MSK17 harbored gene groups associated with directed motility, with many gene groups encoding structural elements of the flagellum, chemotaxis machinery, and associated signaling cascades. In contrast, strains derived from MSK13 and MSK9 lacked many gene groups encoding components of motility and instead contained gene groups associated with the presence of phage—phage plasmid primase activity, DNA methyltransferase activity, and type I restriction modification. Strains from MSK16 were unique, and these strains harbored a subset of gene groups associated with motility but also several gene groups associated with the presence of phage. Collectively, the pattern of gene group presence/absence defined by the Spectral Tree distinguished *E. rectale* strains hierarchically. Strains from MSK22 and MSK17 were more like each other than strains from MSK16, MSK13, and MSK9, and strains from MSK13 and MSK9 were more similar than strains from MSK16.

The statistically deduced patterns of gene group presence/absence motivated testing *E. rectale* isolates from these donors

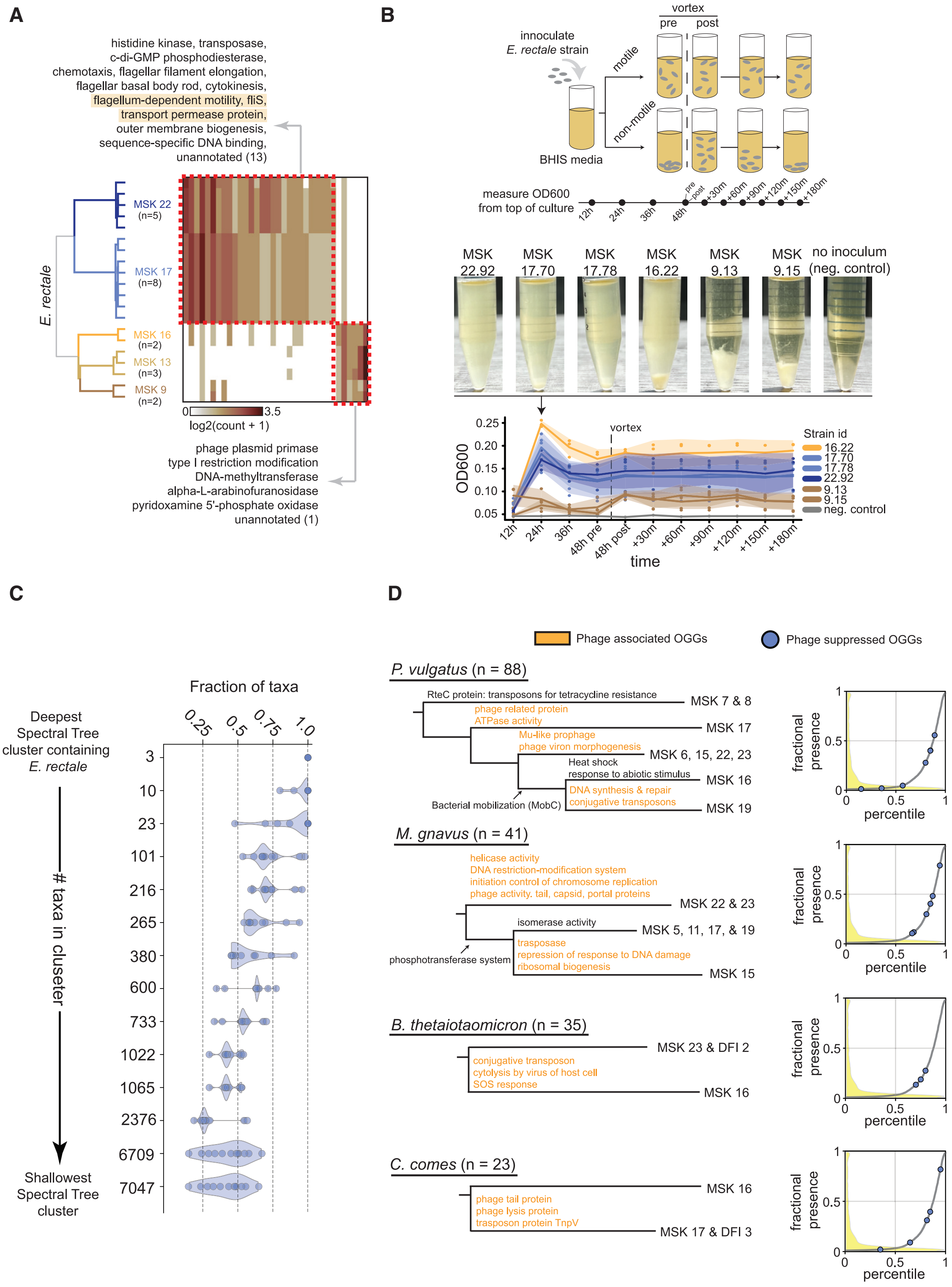


Figure 5. Functional and evolutionary characterization of subspecies phylogeny

(A) Clusters of *Eubacterium rectale* strains from the Spectral Tree (dendrogram). Branches are colored by strain cluster and are labeled by the donor from which they were isolated (MSK indicates Memorial Sloan Kettering Hospital). Number in parenthesis below each donor is number of strains. Heatmap shows gene groups that are significantly differentially abundant between strains defined by the Spectral Tree. Functional annotations of gene groups defining each cluster (red boxes) are shown in text. Highlighted annotations reflect gene groups shared among strains from MSK22, MSK17, and MSK16.

(legend continued on next page)

for their motility. Isolates were tested for their ability to swim in BHIS media (STAR Methods). Six strains, three from each major cluster in Figure 5A, were grown anaerobically for 48 h, vortexed, then observed for 180 min (Figure 5B, top). After 24 h of growth, cultures inoculated from strains of MSK22 and MSK17 were uniformly turbid, illustrating the robust motility of *E. rectale* strains, while those inoculated with strains from MSK9 exhibited a large pellet with clear inoculum. The culture inoculated with strains derived from MSK16 exhibited a phenotype following their pattern of OGG presence/absence—pellet formation with uniform turbidity—illustrating the compromised ability to swim, likely due to the absence of the basal body and other key flagellar and motility components (Figure 5B, middle). The OD_{600} measurements of each culture after vortex were in accord with the phenotypes expected from the pattern of gene group presence/absence for each strain (Figure 5B, lower). Findings in liquid media were also consistent with motility tests performed in solid agar (Figure S12). These results demonstrated that the subspecies phylogeny among *E. rectale* strains inferred from the Spectral Tree manifest as biologically significant differences.

A previously published analysis of *E. rectale* strains demonstrated that a majority of the clade contained motility genes, excepting a single European subspecies, thereby illustrating that motility is a well-conserved trait among *E. rectale*.[38] Thus, our result suggested that subspecies phylogeny associated with phage infection may correlate with strain differences in well-conserved areas of bacterial genomes. We examined the conservation pattern of the 12 annotated gene groups that were absent in *E. rectale* strains isolated from donors MSK16, MSK13, and MSK9 but present in strains isolated from donors MSK22 and MSK17 across the entire Spectral Tree. We found that in the phylogenetic local vicinity of *E. rectale*, the 12 gene groups were well conserved, found in 100% of strains. As we expanded from this vicinity and progressively included more phylogenetically distant bacteria, we found that the 12 gene groups maintained their high conservation, spanning a fractional presence of 20% to greater than 50% across all 7,047 bacteria within UniProt (Figure 5C). These results highlighted that phage-related differences among strains associated with variation among highly conserved *E. rectale* genes.

We then performed a more systematic analysis, focusing on five species outside of *E. rectale* that were represented by more than 20 strain-level variants where differences among gene groups were significant with respect to effect size (log-fold-change greater than 1). These species were *B. uniformis*, *Phocaeicola vulgatus*, *M. gnavus*, *Bacteroides thetaiotaomicron*, and *Coproccocus comes*, comprising 214 strains in total. We found that the most conserved gene groups defining subspecies phylogeny for all species were related to phage physiology (Figure 5D, left). Other features included gene groups related to horizontal gene-transfer and inter-cellular competition, among many other annotations (Table S5). These results were consistent with previous metagenomic-based analyses of subspecies variation in human gut microbiomes illustrating the importance of phage in mediating strain-level variation.[12] We also found that the presence of phage elements correlated with the absence of gene groups that are phylogenetically conserved. We term "phage-suppressed" OGGs as gene groups whose absence was shared with the presence of phage-related gene groups. These groups of OGGs were lost in coordination with the incorporation of phage genomic elements. Across all species that were analyzed, the phage-suppressed OGGs were predominantly within the top half of gene groups ranked by fractional abundance across all 10,177 OGGs defining bacterial proteomes in UniProt. Additionally, several phage-suppressed groups were present in greater than 20% and up to 80% of all taxa in UniProt (Figure 5D, right), illustrating their broad conservation across the kingdom Bacteria. These results show that subspecies phylogeny is markedly associated with a shared history of phage exposure among groups of donors and manifests as functionally relevant changes in clusters of strains due to variation among conserved portions of bacterial genomes. Thus, the origin of subspecies phylogeny in our strain bank was found to be primarily environmentally driven.

Putting together the results of Figures 4 and 5 illustrated that the Spectral Tree resolved structured phylogeny below the level of species in a functionally and evolutionarily relevant manner. Our findings therefore highlight that the Spectral Tree is a more complete phylogenetic description of bacterial strains, motivating using the Spectral Tree to explore genotype-phenotype relationships.

Using the Spectral Tree to understand genotype-metabolic relationships

We next tested whether the Spectral Tree could be used to relate the genotype of individual strains with their metabolism—an important phenotype within the context of the gut ecosystem. To address this idea, we studied species where there were at least 20 representative strains in our strain bank for statistical power. In total, this amounted to 356 strains across 11 species. Instead of describing each strain by their genome—a standard approach in evaluating genotype-phenotype relationships—we coarse-grained the description of strains to their branching pattern in the Spectral Tree. Since strains are linked in the

(B) Evaluating motility of *E. rectale* strains derived from different donors. BHIS media is inoculated with strains, grown for 48 h, vortexed, then observed for 180 min. OD_{600} measurements are taken from the top of the culture. Pictures show cultures of six different strains—three from MSK22 and MSK17 and three from MSK16 and MSK9—and a negative control of media alone after 24 h of culture. OD_{600} (y axis) versus time for each strain in triplicate is shown. Solid lines are average OD_{600} value, contours reflect one standard deviation from average OD_{600} value.

(C) The fraction of taxa (x axis) containing the 12 annotated OGGs (circles) absent in MSK16, MSK13, and MSK9 out of all taxa within a given cluster in the Spectral Tree (y axis). y axis is ordered from the deepest cluster containing the reference *E. rectale* proteome (top) to the shallowest cluster (bottom).

(D) Left: Spectral Tree for given species. Leaves are labeled by donors from which strains were collected, and number of strains collected for each species indicated in parenthesis. Text along branches indicate functional annotation of significantly differentially abundant OGGs between daughter clusters. Orange text indicates annotations associated with phage presence, and black text along daughter cluster indicates functional annotations of OGGs that are absent termed "phage-suppressed" OGGs. Right: all 10,177 OGGs are ordered by their percentile rank of fractional presence in the UniProt database (x axis) and plotted against their fractional presence (y axis) (gray distribution). The density of OGGs for a particular percentile rank is shown in the yellow distribution. Phage-suppressed OGGs—OGGs, which are observed in mutual exclusion of phage-related OGGs—for each species are plotted along the gray distribution in blue circles.

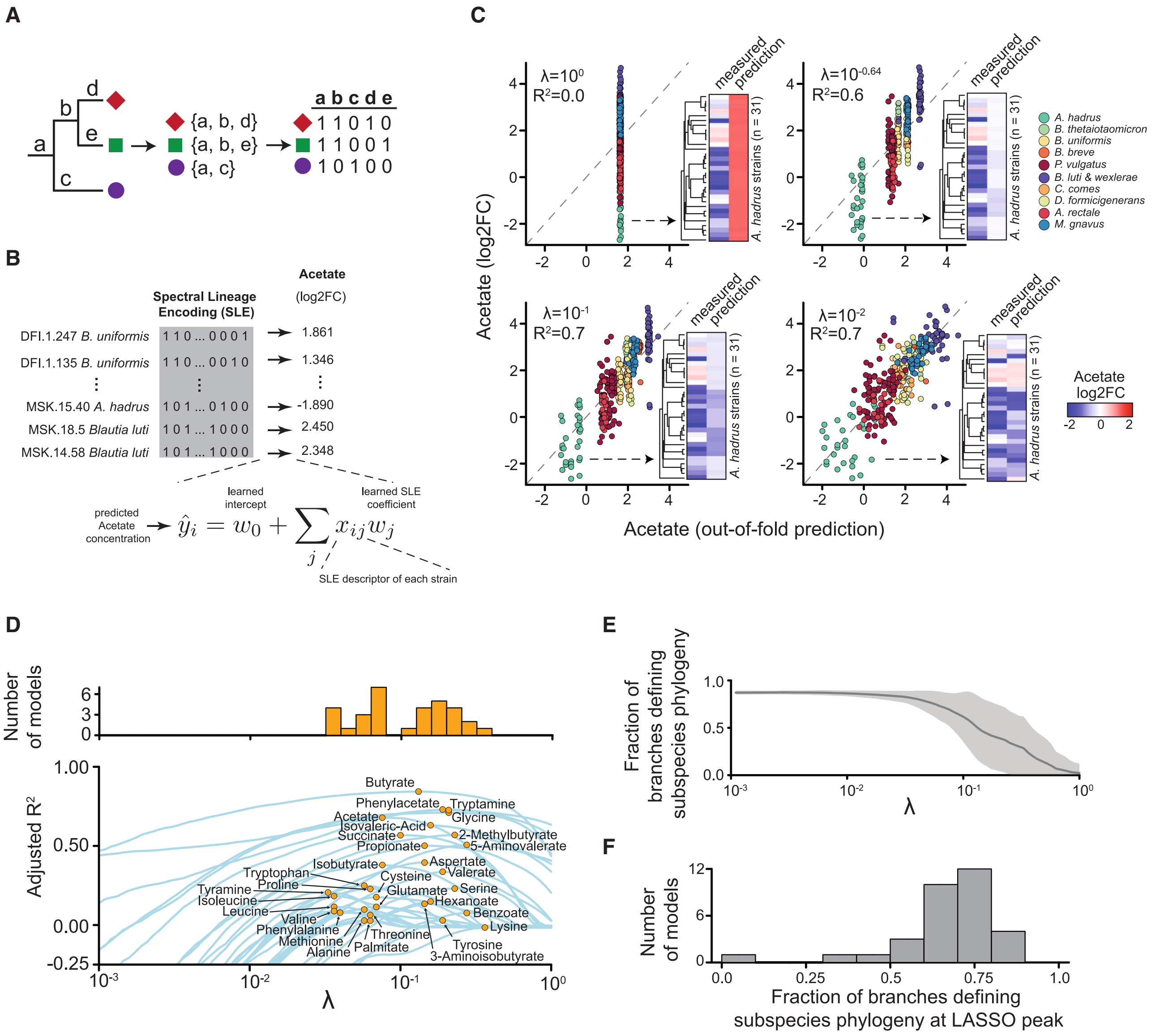


Figure 6. Using the Spectral Tree to relate strain genotype with metabolic capacity

(A) Workflow for defining SLEs for taxa.

(B) Schematic for training LASSO model on SLE designation to predict metabolite concentration, and $\log_2$ fold-change ($\log_2$FC) of acetate concentration relative to a standard in blank media without bacterial culture shown as an example. Resulting models are termed "SLE-LASSO models."

(C) Predicted relative concentration for acetate (x axis) versus measured relative acetate concentration (y axis) for 356 strains spanning different species across decreasing values of LASSO penalty values (λ).

(D) (Bottom) Value of LASSO penalty term (λ, x axis) versus predictive capacity of SLE-LASSO model adjusted by sparsity of model (adjusted r^2, y axis) for each metabolite (blue curves). Solid yellow dots signify the peak predictive capacity of an SLE-LASSO model for a given metabolite. (Top) Number of models with peak predictive capacity (y axis) versus value of LASSO penalty term (λ, x axis).

(E) Value of LASSO penalty term (x axis) versus the fraction of branches in the Spectral Tree that distinguish subspecies phylogeny used in the SLE-LASSO model for a metabolite. Gray distribution reflects the collection of SLE models across all metabolites, and black solid line is the average fraction.

(F) Number of models (y axis) versus the fraction of branches in the Spectral Tree defining subspecies phylogeny within the best-performing SLE-LASSO models (yellow dots in D).

Spectral Tree by a common root, the branching pattern of each strain could be used as a unique "barcode" of statistically inferred evolutionary lineage. We therefore termed this barcode a "spectral lineage encoding" (SLE) (Figure 6A). Next, we reasoned that training statistical models on patterns of SLE descriptions of strains would be a way to test whether phylogenetic information could be directly related to metabolic phenotype. Thus, we trained LASSO models that used the SLE for each bacterial strain as input and the relative difference in metabolite concentrations for all metabolites that we profiled as output

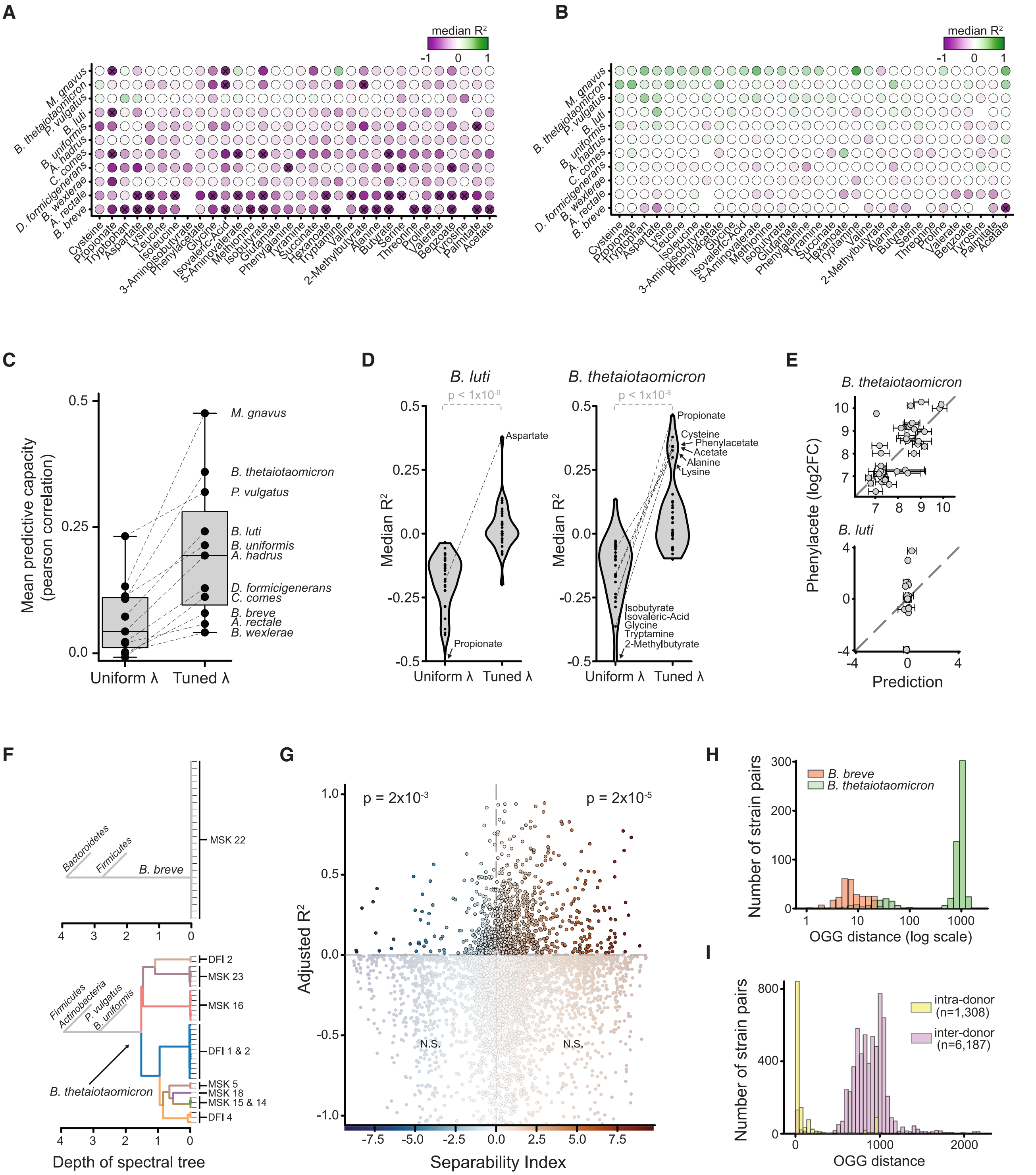


Figure 7. Inter-donor variation between strain-level variants drives predictive capacity of SLE-LASSO models of strain metabolism

(A and B) The predictive capacity of SLE-LASSO models when considering a uniform LASSO penalty value across all species for each metabolite (A) or a LASSO penalty term tuned to optimize the predictive capacity for each species/metabolite pair (B).

(C) Mean predictive capacity of SLE-LASSO models in (A) (“uniform λ”) or (B) (“tuned λ”) averaged across metabolites for each species.

(D) Median predictive capacity for SLE-LASSO models of *Blautia luti* (left) and *Bacteroides thetaiotaomicron* (right) strains for each metabolite (dots) where models were trained using either a uniform LASSO penalty term across all species for a given metabolite (“uniform λ) or a LASSO penalty term tuned for the specific species/metabolite pair (tuned λ). *p* values between violin plots for each species are computed by a Mann-Whitney rank-order test.

(legend continued on next page)

(Figure 6B). To ensure an out-of-sample prediction for all strains, we computed the spectral distance of a randomly chosen 75% of the 356 strains within the Spectral Tree as a training set. Next, we assigned SLEs to all strains in the training set, trained a LASSO model, and validated the LASSO model on the remaining 25% of strains. We then repeated these steps four times across five different repartitions of the dataset (Figure S13) (STAR Methods).

Generally, in training LASSO models, a penalty term (also known as a "regularization parameter") is used to constrain the number of coefficients in the resulting model. The larger the penalty term, the fewer coefficients are used for the predictive capacity of the model. In other words, the penalty term is used to "coarsen" the model. We found that as the LASSO penalty term was reduced, the SLE-LASSO models became progressively better at predicting bacterial metabolism at finer scales of phylogeny. For instance, when training models to predict acetate metabolism, using a penalty term of 1 collapsed all predictions onto the acetate concentration averaged across all strains (Figure 6C, top left). As the penalty term was decreased, the range of predicted acetate levels increased (Figure 6C, top right and bottom left). At a penalty term value of 10^{-2}, we found that the predictive capacity of the SLE-LASSO models differentiated strain-level differences between acetate consumers and producers (Figure 6C, bottom right and inset). Thus, these results motivated the idea that lowering the LASSO penalty terms progressively incorporated deeper branches of the Spectral Tree, thereby allowing SLE-LASSO models to increasingly consider strain-level genomic differences. Consistent with this result, we also observed that if the training set to the SLE-LASSO models was missing coarse phylogenetic structure, e.g., branches of entire species that define sets of strains, the SLE-LASSO models were unable to predict the average metabolic capacity of the species irrespective of the penalty value (Figure S14). These results motivated the hypothesis that the LASSO penalty term may be a tuning parameter that is directly related to phylogenetic structure as opposed to a hyperparameter that restructures the genomic neighborhoods of strains in a metabolically aware manner akin to a deep-neural network. We therefore next investigated the relationship between (1) the LASSO penalty term, (2) the predictive capacity of the SLE-LASSO models, and (3) the scale of phylogeny being considered in the models for all metabolites.

For each metabolite, we trained and validated SLE-LASSO models across a range of penalty terms. We found that when adjusted for model parsimony, the best predictive capacities for our models (calculated as adjusted r^2 values) occurred between a penalty term value of $10^{-0.5}$ and $10^{-1.5}$ (Figure 6D; Table S6). Importantly, we also found that as the penalty term continued to increase beyond these values, the predictive capacity of our models decreased (Figure S15A). These observations demonstrated that there was a range of optimal penalty term values that (1) balanced the degree to which models should be coarse-grained for achieving optimal predictive capacity of metabolites and (2) prevented overfitting or underfitting of models relative to the training set used to train the models. To relate this result with scale of phylogeny being considered in the models, we investigated how defining the optimal penalty value for our SLE-LASSO models affected the degree to which subspecies phylogeny was being considered by the models. To address this, we quantified the number of Spectral Tree branches defining subspecies phylogenetic clusters that were being considered by the model at the penalty term associated with the peak predictive capacity for each metabolite. We found that as the penalty term decreased, the fraction of Spectral Tree branches defining subspecies phylogeny being used in the model increased (Figure 6E). We found that nearly all peak predictive SLE-LASSO models incorporated Spectral Tree branches that defined clusters of subspecies phylogeny with non-zero coefficients (Figures 6F and S15B). As an example, interrogating the SLE-LASSO model associated with the peak predictive capacity for acetate illustrated the presence of several non-zero LASSO coefficients derived from considering subspecies phylogenetic structure in the Spectral Tree (Figure S16).

These results demonstrated that by resolving subspecies phylogeny, the Spectral Tree can enable learning genotype-metabolic relationships for individual strains. However, we noted that the range of predicting strain-level metabolic capacity per our SLE-LASSO models was large (Figure S17). For instance, butyrate could be predicted up to an adjusted R^2 value of 0.84 while many amino acids could be predicted only up to an adjusted R^2 value of <0.2 (Figure 6D). We therefore sought to better understand why the predictive capacities of certain metabolites were markedly better than others.

First, we found that SLE-LASSO models using the penalty parameter associated with the peak predictive capacity determined by considering all species exhibited a poor capacity overall to predict the metabolism of strains within specific species (Figure 7A; Table S7A). Across all metabolites, a majority of predictive capacities for a given species were close to or less than zero, highlighting the paucity of predictive power of our SLE-LASSO models on a per-species basis. In contrast, we found that if we treated the LASSO penalty value as a hyperparameter and tuned each species-by-metabolite relationship separately, the predictive capacity of strain metabolism within species increased for all species on average (Figures 7B and 7C; Tables S7B and S7C). Moreover, we found that the resulting distribution of penalty terms tuned for each species/metabolite pair

(E) Relative concentration of phenylacetate (y axis) versus predicted relative concentration (x axis) where predictions are made by SLE-LASSO models for which the penalty term is tuned to the species metabolite/pair for strains of *B. thetaiotaomicron* (top) and *B. luti* (bottom).

(F) Architecture of Spectral Tree for strains of *Bifidobacterium breve* (top) and *B. thetaiotaomicron* (bottom).

(G) Adjusted R^2 values for all 7,040 repartitioned SLE-LASSO models from (B) (y axis) versus the separability index (x axis) measured by metabolic variation between donors (see STAR Methods for definition). *p* values in quadrants reflect statistical significance of enrichment or depletion by Fisher's exact test using Bonferroni correction (see Figure S20 for workflow).

(H and I) Number of strain pairs (y axis) versus distance between strain pairs based on the OGG profile of a strain (x axis). (H) Distributions shown for strains belonging to *B. breve* (orange distribution) and *B. thetaiotaomicron* (green distribution). (I) Distributions shown for all pairs of strains sharing the same donor (intra-donor, yellow distribution) or different donors ("inter-donor," purple distribution).

included a substantial fraction of Spectral Tree branches that defined strain clusters at the scale of subspecies phylogeny (Figure S18). Consistent with this finding, setting the coefficients of SLE-LASSO models corresponding to subspecies phylogeny to zero significantly reduced the predictive capacity of the models (Figure S19). This result therefore illustrated that merely knowing the species designation of strains was insufficient for predicting their metabolic capacity; rather, it was necessary to incorporate subspecies phylogenetic information even when treating each species/metabolite pair separately with respect to the LASSO penalty terms.

We next found that the increased capacity to predict strain-level metabolic phenotype was dependent on the specific species/metabolite pair and not driven by a stereotyped class of metabolites. For example, we compared the SLE-LASSO models of *Blautia luti* and *Bacteroides thetaiotaomicron* for all metabolites using either a uniform penalty term across all species or a penalty term tuned for each species/metabolite pair. As expected from the results shown in Figure 7B, we found that the predictive capacities for both species significantly increased (Figure 7D). However, the metabolites with the highest predictive capacities were different for each species—the predictive capacity of aspartate metabolism for strains of *B. luti* exhibited an R^2 of greater than 0.25, while metabolites associated with an equivalent predictive capacity for strains of *B. thetaiotaomicron* were propionate, cysteine, phenylacetate, acetate, alanine, and lysine (Figures 7D and 7E).

Collectively, these findings illustrated two results. First, by treating the LASSO penalty term as a hyperparameter that can be tuned for a given species and metabolite of interest, strain-level metabolic capacity can be learned in specific cases of species/metabolite pairings. Second, the ability to learn strain-level metabolic capacity may not be generally possible across species or metabolites, suggesting either the need for increased sampling or that the biology underlying the metabolic capacity of species may originate from variation outside of genomic information.

However, the results described above also motivated a key question: if the ability to predict strain-level metabolic capacity is dependent on the specific species/metabolite pairing, is there any measure or descriptor that could inform whether the metabolic capacity of a strain can be learned from SLE-based statistical models? To address this question, we turned to two species—*Bifidobacterium breve* and *B. thetaiotaomicron*—as a case study because the predictive capacities of SLE-based LASSO models are uniformly poor for *B. breve* but are predictive for certain metabolites for *B. thetaiotaomicron*. We found that all *B. breve* strains were collected from the same donor, resulting in a “flat” Spectral Tree architecture below the level of species (Figure 7F, top). In contrast, strains of *B. thetaiotaomicron* were collected from different donors and manifest in a structured Spectral Tree architecture below the level of species (Figure 7F, bottom).

This result motivated the hypothesis that strain-level variants collected from different donors are genetically more diverse than strain-level variants collected from the same donor, thereby introducing more genetic variation that can be captured by the Spectral Tree to define subspecies phylogeny and therefore learn better SLE-based models of strain metabolic capacity. We tested this hypothesis by investigating SLE-LASSO models from Figure 7B and interrogated the role of metabolic variation within a single donor versus between different donors on influencing the resulting predictive capacity of the models. The predictive capacity of each strain/metabolite pair shown in Figure 7B was the median predictive capacity of 20 separate models, each trained on a different repartitioning of the dataset. Thus, the total number of models reflected in Figure 7B was 7,040 (20 repartitions, 32 metabolites, 11 species). First, we stratified the test set for each metabolite-species-repartition combination by donor. Second, we calculated the mean and standard deviation of relative metabolite concentrations for strains within each donor. Third, we defined the standard deviation of the means as inter-donor metabolic variation, and we defined the mean of the standard deviations as intra-donor metabolic variation. Thus, the ratio between inter- and intra-donor metabolic variation was a measure of metabolic separability by donor (Figure S20) (STAR Methods). We term the $\log_2$ fold-change of this ratio the “separability index.” A separability index of greater than 1 indicated that relative metabolite concentration was separable by donor, and a separability index of less than 1 indicated metabolic variability across donors was not separable. Using this metric, we found there to be a statistically significant enrichment for higher predictive capacity in SLE-LASSO models where the relative metabolite concentrations in the test set were separable by donor (Figure 7G; Table S8). This finding suggested that by collecting strains across different donors, we would increase the likelihood of introducing metabolic variability manifest through differences in strain genomes, motivating comparing the genomic composition of strains collected from different donors and a single donor. Analyzing the OGG content of strains, we found that inter-donor strain-level variation was significantly greater with respect to genomic diversity relative to intra-donor strain-level variation (Figures 7H and 7I). Together, these results illustrated that one measure indicative of the capacity to learn strain-level metabolic qualities from genomes is the presence of subspecies-level structure in the Spectral Tree—a property we found to be more likely when sampling strains of the same species from different donors. We discuss the implications of this finding with respect to informative sampling of bacterial strains in the discussion.

Considerations for creating and using Spectral Trees

We outline two sets of considerations regarding our work. The first is with respect to elucidating subspecies phylogeny using the Spectral Tree. The second is with respect to using the Spectral Tree to relate strain genotype with phenotype.

With regard to resolving subspecies phylogeny within our strain bank, we note two important aspects that enabled our result. First, the Spectral Tree created from the UniProt database incorporated a wide breadth of diversity, encompassing bacterial proteomes across a range of econiches, some of which included the econiche of the human gut. Moreover, we found that this diversity was crucial for constructing an accurate Bayesian prior to contextualize variation among strains in our strain bank. However, we note that we currently do not have a quantitative metric for determining the extent of econiche diversity that is necessary to include in order to resolve subspecies phylogeny. As such, a limitation of our approach is being unable to quantitatively define the extent of background diversity needed for resolving subspecies phylogeny within a strain bank. Second, by using OGGs as the set of features to describe a proteome, we captured markedly more genome space than canonical descriptions of strains based on 16S or sets of gene markers. The rationale behind expanding genomic descriptors of bacterial phylogeny from 16S to sets of marker genes, and ultimately bac120—a set of 120 genes that define phylogenetic relationships between bacteria—was to balance "signal-to-noise": capture enough genomic variation to robustly define phylogeny while avoiding saturation of genomic variation with highly fluctuant information. In this sense, we note that though the Spectral Tree uses the whole proteome as input, spectral decomposition organizes the variation into hierarchical scales of signal such that unstructured variation (e.g., statistical patterns that cannot be distinguished from noise) is placed at the bottom of the eigenspectrum. The rationale for using a more high-content feature set for describing strain genomes is not new as others have employed comparative methods across bacteria at the level of amino acid resolution or in defining bacterial "pangenomes"—conserved genomic elements across phylogenetically similar strains.[39,40] A more detailed explanation comparing pangenome analysis to the construction of Spectral Trees can be found in supplemental experimental procedures Section 3 (Figure S21). These approaches suggest that there may be a degeneracy of different sets of features that effectively access information across the whole bacterial proteome. The practical implication of these two considerations is that when applying our framework to new strain banks, it is important to ensure that (1) the Spectral Tree is created from a diverse set of proteomes and (2) that the proteomes are described in a sufficiently high-content manner.

With regard to using the Spectral Trees to relate bacterial genotype with phenotype, we note three caveats to consider. First, the SLE-based approach we developed presumes a genetic basis for phenotype that can be accurately captured in OGGs. It is possible that the phenotype of interest may be reflected in other descriptions of genetic information—i.e., amino acid changes within protein sequences belonging to the same OGGs and insertion/deletion ("indels") mutations within genes—or in non-genetic mechanisms of action like transcriptional changes, interactions with other microbes, or interactions with the environment. In either of these cases, our framework will not produce a predictive model of phenotype. Second, while the set of OGGs differentiating strains from each other was associated with strain-level metabolism, understanding the biological mechanism underlying our results is immensely challenging due to the unannotated nature of OGGs. The analysis we performed identifying OGGs that separated layers of the Spectral Tree showed that while OGGs differentiating coarse phylogeny (phylum to species) were annotated to an extent above 80%, greater than 40% of OGGs differentiating individual strains were unannotated (Figure S22A). Moreover, metabolic variability captured by the Spectral Tree is associated with OGGs that are annotated by metabolic functions at a phylum-to-species level but are broadly unannotated at the level of subspecies phylogeny (Figures S22B and S22C). Therefore, we note that validating the OGGs responsible for determining strain-level metabolic capacity or other phenotypes within species will first require performing precise experiments to functionally annotate the candidate OGGs and then understand how their variation affects bacterial metabolism. As tools for genetic manipulation in bacteria outside of well-studied model organisms are limited in their development, we anticipate this remaining a significant challenge for the immediate future. Third, the capacity to use the SLE-LASSO models to predict metabolic phenotypes of subspecies phylogeny is, by definition, dependent on capturing phylogenetic diversity in the training set for the model. The reason for this is because the SLE-LASSO model is a direct representation of the statistical geometry of genomic variation across bacteria. Our results have shown that this geometry reflects phylogenetic scales of organization. Therefore, if the training set is missing a portion of coarse phylogenetic structure (e.g., an entire species), the resulting SLE-LASSO model will not be able to predict strain-level metabolic capacities for the missing species. Of note, the SLE-LASSO models are a fundamentally different statistical architecture than other types of models that could be used like artificial neural networks (ANNs), where the statistical geometry of variation, and therefore phylogenetic relationships, are contorted to be phenotypically aware.

DISCUSSION

The importance of individual strains in mediating gut microbiome function requires new frameworks for their description beyond merely taxonomic definitions. Here, we showed that co-evolutionary patterns learned from a large diversity of strains across the bacterial kingdom creates a natural, data-driven, and useful description of gut commensal strains, revealing the existence of phylogenetic structure below the level of species. Importantly, our findings demonstrate how leveraging biological diversity reflective of many diverse and unrelated environments can expose constraints on genomic variation within a single environment. As our framework is not specific to gut bacteria but can be applied to strains isolated from any environment, we pose that the construct we have developed—the SLE—may be a generally useful schema for describing and studying bacterial strains.

The intra- and interpersonal variation in the structure of human gut microbiomes has been extensively described.[41–43] The degree to which this variation reproducibly derives from external factors has remained a subject of discussion with recent studies attempting to control for environment—like diet or spatial geography—to "normalize" structural changes observed in human cohort studies.[44–46] Our data suggest that a history of phage infection among hosts can lead to structured, non-random microbiome changes between groups of humans that manifest in subspecies phylogeny. While we demonstrated how these changes lead to different behaviors at the scale of individual bacteria, the functional consequences of such strain-level variation at the scale of the whole microbiome remain to be characterized. Though a majority of the phage-suppressed OGGs we identified were phylogenetically conserved, the strains nevertheless persisted in the gut microbiome of donors. This suggests that perhaps changes in conserved genomic areas within individual bacteria can be tolerated without a substantial fitness decrease when considered within the context of the entire gut ecosystem. The recent shift toward genomic analysis of bacteria within the context of whole microbial ecosystems will enable a better definition for a “null hypothesis” of genomic constraint within individual strains.

From a practical perspective of learning about the metabolic capacity of individual strains, it has been previously argued that because strain-level genomic variation does not obviously map to strain-level metabolic variation, it is necessary to metabolically profile each and every new strain that is collected.[20] Our results suggest a substantially different point of view. As more genetically diverse bacterial strains are collected, sequenced, and metabolically phenotyped, including new metabolites that are discovered to be important, the constructs developed here—the Spectral Tree and SLE-based predictive models—could be used to learn genotype-metabolic relationships of strain-level variants. Indeed, as our results demonstrate, achieving a reasonable predictive capacity even for a subset of metabolites required tuning our statistical models in a manner specific to the species-metabolite pairing. However, our findings also showed that increasing the genetic diversity of subspecies phylogeny is directly related to creating predictive statistical models of metabolic capacity. Thus, our results highlight the predictive power in having a strain bank comprising diverse subspecies phylogeny. How can this practically be achieved? As we showed that genetic diversity in subspecies phylogeny originates from the econiche of individual donors (see Figures 4C and 7F), we pose that a useful sampling strategy is constructing strain banks across a broad set of donors rather than deeply sampling strains from individual donors. Our data suggest that this approach to sampling—shallow sampling across many donors—will introduce the necessary scale of genomic variation for learning genotype-phenotype relationships among strain-level variants of the same species. We acknowledge that an important caveat is that strain metabolism can change as a function of culture conditions; therefore, it will be important in the future to test whether coordinated changes in bacterial metabolism across culture conditions also follow co-evolutionary patterns as described here. However, because the Spectral Tree is an object capable of incorporating new sequences and as there are many ongoing efforts to understand bacterial genotype-phenotype relationships at the scale of individual strains, the Spectral Tree could be a unifying dynamic framework for performing comparative phylogenomics for arbitrary phenotypes of interest.

What fundamental properties underlie the utility of describing strains by their co-evolutionary signature? Unlike engineered systems, existing or “extant” biological systems arise from ancestors through the evolutionary process.[47–51] Therefore, understanding how patterns of genetic interactions encode behaviors is inextricably intertwined with defining commonalities and divergences in molecular structure.[30,52–57] Current statistical strategies for parsing differences among genomes involve so-called “factorization” approaches that discover low-dimensional representations of high-dimensional patterns of variation. Indeed, the era of biological big data has seen an explosion in the use of factorization methods.[58] Such approaches are predicated on a key assumption: the systems being interrogated are unrelated to each other. For evolved systems, ancestral relatedness violates the assumption of system independence, demanding a new formalism for comparative efforts. We reason that the SLE is a useful descriptor of bacteria because it embeds hierarchical scales of relatedness across the evolutionary record, simultaneously capturing both broad phylogenetic differences and fine-grained differences within species. The hierarchical nature of the SLE as a descriptor therefore distinguishes variation arising from ancient phylogenetic sources from functional differences reflecting recent adaptations (i.e., to individual human hosts). This capacity to separate sources of variance is key for creating accurate predictive models of biological behavior from genome content.[59] We anticipate that our approach is unlikely to be the only applicable framework given the recent development, implementation, and success of large language models (LLMs) in characterizing evolutionary relationships among complex biological systems.[60,61] However, our findings show that creating statistical representations of the evolutionary record may lay an interpretable foundation for understanding and predicting idiosyncrasies of individual biological systems that deviate from broad phylogenetic trends. Future studies applying the concepts developed here to other evolved systems will test this idea.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Arjun S. Raman (araman@bsd.uchicago.edu).

Materials availability

Strains of the commensal strain bank are available upon request.

Data and code availability

- Sequencing data for all strains used in this paper are available on NCBI: PRJNA737800. All raw data from metabolomics can be downloaded from Metabolights: MTBLS7771.
- All analysis code used in this study for analysis was made in Julia. Some code for plotting was made in R. All code used for genome sequencing and metabolomic processing was made in Python. All code, along with annotations and stepwise instructions, are available for download via a github repository: https://github.com/aramanlab/Doran_etal_2023. (https://doi.org/10.5281/zenodo.13244426)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank S. Kuehn, M. Mani, D. Pincus, A. Murugan, J. Gordon, S. Oakes, A. Drummond, O. Rivoire, and R. Ranganathan for helpful discussions. This work was supported by NIH grant RM35GM146702 and the Duchossois Family Institute at the University of Chicago.

AUTHOR CONTRIBUTIONS

B.A.D. and A.S.R. designed this study and conceived of the approach taken. B.A.D. and R.Y.C. conceptualized “spectral distance” as a quantitative metric. H.G. and B.A.D. performed motility experiments of *E. rectale* strains. B.B., A.S., and H.L. oversaw the collection, isolation, sequencing, and bioinformatic analysis of the CSB. V.B. conceptualized and wrote the code for evaluating statistically significant differences in feature abundances between daughters of Spectral Tree. A.S. aided in the execution and supervision of metabolomic profiling on strains. E.G.P. oversaw and supervised the collection, isolation, sequencing, and metabolomic profiling of strains. B.A.D. wrote all code and conducted all analysis. B.A.D. and A.S.R. wrote the paper.

DECLARATION OF INTERESTS

E.G.P. serves on the advisory board of Diversigen, has received speaker honoraria from Bristol Myers Squibb, Celgene, Seres Therapeutics, MedImmune, Novartis, and Ferring Pharmaceuticals; is an inventor on patent applications WPO2015179437A1, entitled "Methods and compositions for reducing *Clostridium difficile* infection," and WO2017091753A1, entitled "Methods and compositions for reducing vancomycin-resistant enterococci infection or colonization"; and holds patents that receive royalties from Seres Therapeutics Inc.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Genome sequencing of commensal strain bank
  - ○ Metabolic profiling of strains
  - ○ Motility assay for E. rectale isolates
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Phylogenetic trees of strain bank
  - ○ Spectral distance and spectral groups
  - ○ Measuring the accuracy of Spectral Trees
  - ○ Mutual Information (MI) calculation
  - ○ Recreating the alignment
  - ○ Creating a Spectral Tree across UniProt
  - ○ MI between Spectral Tree and phylogeny
  - ○ Projecting CSB into the Spectral Tree
  - ○ SLE LASSO model training and validation
  - ○ Separability of donor metabolic variation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2024.12.008.

Received: December 26, 2023
Revised: February 16, 2024
Accepted: December 18, 2024
Published: January 17, 2025

REFERENCES

1. Sunagawa, S., Acinas, S.G., Bork, P., Bowler, C., Tara Oceans Coordinators, Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., et al. (2020). Tara Oceans: towards global ocean ecosystems biology. Nat. Rev. Microbiol. *18*, 428–445. https://doi.org/10.1038/s41579-020-0364-5.

2. Integrative HMP (iHMP) Research Network Consortium (2019). The Integrative Human Microbiome Project. Nature *569*, 641–648. https://doi.org/10.1038/s41586-019-1238-8.

3. Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021). A genomic catalog of Earth's microbiomes. Nat. Biotechnol. *39*, 499–509. https://doi.org/10.1038/s41587-020-0718-6.

4. Compant, S., Samad, A., Faist, H., and Sessitsch, A. (2019). A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. J. Adv. Res. *19*, 29–37. https://doi.org/10.1016/j.jare.2019.03.004.

5. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227. https://doi.org/10.1038/nature11053.

6. Thursby, E., and Juge, N. (2017). Introduction to the human gut microbiota. Biochem. J. *474*, 1823–1836. https://doi.org/10.1042/BCJ20160510.

7. Cho, I., and Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. Nat. Rev. Genet. *13*, 260–270. https://doi.org/10.1038/nrg3182.

8. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

9. Yan, Y., Nguyen, L.H., Franzosa, E.A., and Huttenhower, C. (2020). Strain-level epidemiology of microbial communities and the human microbiome. Genome Med. *12*, 71. https://doi.org/10.1186/s13073-020-00765-y.

10. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638. https://doi.org/10.1101/gr.216242.116.

11. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife *10*, e65088. https://doi.org/10.7554/eLife.65088.

12. Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960. https://doi.org/10.15252/msb.20177589.

13. Barratt, M.J., Nuzhat, S., Ahsan, K., Frese, S.A., Arzamasov, A.A., Sarker, S.A., Islam, M.M., Palit, P., Islam, M.R., Hibberd, M.C., et al. (2022). *Bifidobacterium infantis* treatment promotes weight gain in Bangladeshi infants with severe acute malnutrition. Sci. Transl. Med. *14*, eabk1107. https://doi.org/10.1126/scitranslmed.abk1107.

14. Sela, D.A., Garrido, D., Lerno, L., Wu, S., Tan, K., Eom, H.-J., Joachimiak, A., Lebrilla, C.B., and Mills, D.A. (2012). Bifidobacterium longum subsp. infantis ATCC 15697 α-fucosidases are active on fucosylated human milk oligosaccharides. Appl. Environ. Microbiol. *78*, 795–803. https://doi.org/10.1128/AEM.06762-11.

15. Underwood, M.A., German, J.B., Lebrilla, C.B., and Mills, D.A. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. Pediatr. Res. *77*, 229–235. https://doi.org/10.1038/pr.2014.156.

16. Yang, C., Mogno, I., Contijoch, E.J., Borgerding, J.N., Aggarwala, V., Li, Z., Siu, S., Grasset, E.K., Helmus, D.S., Dubinsky, M.C., et al. (2020). Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides ovatus and Are Modifiable by Gut Microbiota Manipulation. Cell Host Microbe *27*, 467–475.e6. https://doi.org/10.1016/j.chom.2020.01.016.

17. Patnode, M.L., Guruge, J.L., Castillo, J.J., Couture, G.A., Lombard, V., Terrapon, N., Henrissat, B., Lebrilla, C.B., and Gordon, J.I. (2021). Strain-level functional variation in the human gut microbiota based on bacterial binding to artificial food particles. Cell Host Microbe *29*, 664–673.e5. https://doi.org/10.1016/j.chom.2021.01.007.

18. Hall, A.B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., Lagoudas, G.K., Vatanen, T., Fornelos, N., Wilson, R., et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med. *9*, 103. https://doi.org/10.1186/s13073-017-0490-5.

19. Lianou, A., and Koutsoumanis, K.P. (2013). Strain variability of the behavior of foodborne bacterial pathogens: a review. Int. J. Food Microbiol. *167*, 310–321. https://doi.org/10.1016/j.ijfoodmicro.2013.09.016.

20. Han, S., Van Treuren, W., Fischer, C.R., Merrill, B.D., DeFelice, B.C., Sanchez, J.M., Higginbottom, S.K., Guthrie, L., Fall, L.A., Dodd, D., et al. (2021). A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. Nature *595*, 415–420. https://doi.org/10.1038/s41586-021-03707-9.

21. Chen, H., Nwe, P.-K., Yang, Y., Rosen, C.E., Bielecka, A.A., Kuchroo, M., Cline, G.W., Kruse, A.C., Ring, A.M., Crawford, J.M., et al. (2019). A Forward Chemical Genetic Screen Reveals Gut Microbiota Metabolites

That Modulate Host Physiology. Cell *177*, 1217–1231.e18. https://doi.org/10.1016/j.cell.2019.03.036.

22. Sorbara, M.T., Littmann, E.R., Fontana, E., Moody, T.U., Kohout, C.E., Gjonbalaj, M., Eaton, V., Seok, R., Leiner, I.M., and Pamer, E.G. (2020). Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. Cell Host Microbe *28*, 134–146.e4. https://doi.org/10.1016/j.chom.2020.05.005.

23. Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589–594. https://doi.org/10.1016/j.gde.2005.09.006.

24. Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. Genetics *194*, 573–596. https://doi.org/10.1534/genetics.113.151753.

25. R Oaks, J., A Cobb, K., N Minin, V., and D Leaché, A. (2019). Marginal Likelihoods in Phylogenetics: A Review of Methods and Applications. Syst. Biol. *68*, 681–697. https://doi.org/10.1093/sysbio/syz003.

26. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004. https://doi.org/10.1038/nbt.4229.

27. Sanna, S., van Zuydam, N.R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Võsa, U., Mujagic, Z., Masclee, A.A.M., Jonkers, D.M.A.E., Oosting, M., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat. Genet. *51*, 600–605. https://doi.org/10.1038/s41588-019-0350-x.

28. Liu, J., Tan, Y., Cheng, H., Zhang, D., Feng, W., and Peng, C. (2022). Functions of Gut Microbiota Metabolites, Current Status and Future Perspectives. Aging Dis. *13*, 1106–1126. https://doi.org/10.14336/AD.2022.0104.

29. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

30. Zaydman, M.A., Little, A.S., Haro, F., Aksianiuk, V., Buchser, W.J., DiAntonio, A., Gordon, J.I., Milbrandt, J., and Raman, A.S. (2022). Defining hierarchical protein interaction networks from spectral analysis of bacterial proteomes. eLife *11*, e74104. https://doi.org/10.7554/eLife.74104.

31. Sakoparnig, T., Field, C., and van Nimwegen, E. (2021). Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. eLife *10*, e65366. https://doi.org/10.7554/eLife.65366.

32. Magee, A.F., Hilton, S.K., and DeWitt, W.S. (2021). Robustness of Phylogenetic Inference to Model Misspecification Caused by Pairwise Epistasis. Mol. Biol. Evol. *38*, 4603–4615. https://doi.org/10.1093/molbev/msab163.

33. Wigner, E.P. (1967). Random Matrices in Physics. SIAM Rev. *9*, 1–23. https://doi.org/10.1137/1009001.

34. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

35. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

36. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat. Microbiol. *1*, 16048. https://doi.org/10.1038/nmicrobiol.2016.48.

37. Wang, Y., Huang, J.-M., Zhou, Y.-L., Almeida, A., Finn, R.D., Danchin, A., and He, L.-S. (2020). Phylogenomics of expanding uncultured environmental Tenericutes provides insights into their pathogenicity and evolutionary relationship with Bacilli. BMC Genomics *21*, 408. https://doi.org/10.1186/s12864-020-06807-4.

38. Karcher, N., Pasolli, E., Asnicar, F., Huang, K.D., Tett, A., Manara, S., Armanini, F., Bain, D., Duncan, S.H., Louis, P., et al. (2020). Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol. *21*, 138. https://doi.org/10.1186/s13059-020-02042-y.

39. Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. Science *365*, 185–189. https://doi.org/10.1126/science.aaw6718.

40. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

41. Zhao, S., Lieberman, T.D., Poyet, M., Kauffman, K.M., Gibbons, S.M., Groussin, M., Xavier, R.J., and Alm, E.J. (2019). Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe *25*, 656–667.e8. https://doi.org/10.1016/j.chom.2019.03.007.

42. Garud, N.R., Good, B.H., Hallatschek, O., and Pollard, K.S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. *17*, e3000102. https://doi.org/10.1371/journal.pbio.3000102.

43. Yang, Y., Nguyen, M., Khetrapal, V., Sonnert, N.D., Martin, A.L., Chen, H., Kriegel, M.A., and Palm, N.W. (2022). Within-host evolution of a gut pathobiont facilitates liver translocation. Nature *607*, 563–570. https://doi.org/10.1038/s41586-022-04949-x.

44. Gehrig, J.L., Venkatesh, S., Chang, H.-W., Hibberd, M.C., Kung, V.L., Cheng, J., Chen, R.Y., Subramanian, S., Cowardin, C.A., Meier, M.F., et al. (2019). Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. Science *365*, eaau4732. https://doi.org/10.1126/science.aau4732.

45. Delannoy-Bruno, O., Desai, C., Raman, A.S., Chen, R.Y., Hibberd, M.C., Cheng, J., Han, N., Castillo, J.J., Couture, G., Lebrilla, C.B., et al. (2021). Evaluating microbiome-directed fibre snacks in gnotobiotic mice and humans. Nature *595*, 91–95. https://doi.org/10.1038/s41586-021-03671-4.

46. Guthrie, L., Spencer, S.P., Perelman, D., Van Treuren, W., Han, S., Yu, F.B., Sonnenburg, E.D., Fischbach, M.A., Meyer, T.W., and Sonnenburg, J.L. (2022). Impact of a 7-day homogeneous diet on interpersonal variation in human gut microbiomes and metabolomes. Cell Host Microbe *30*, 863–874.e4. https://doi.org/10.1016/j.chom.2022.05.003.

47. Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA *74*, 5088–5090. https://doi.org/10.1073/pnas.74.11.5088.

48. Felsenstein, J. (1985). Phylogenies and the Comparative Method. Am. Nat. *125*, 1–15. https://doi.org/10.1086/284325.

49. Woese, C. (1998). The universal ancestor. Proc. Natl. Acad. Sci. USA *95*, 6854–6859. https://doi.org/10.1073/pnas.95.12.6854.

50. Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314. https://doi.org/10.1038/nrg3186.

51. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. Nat. Rev. Genet. *21*, 428–444. https://doi.org/10.1038/s41576-020-0233-0.

52. Qin, C., and Colwell, L.J. (2018). Power law tails in phylogenetic systems. Proc. Natl. Acad. Sci. USA *115*, 690–695. https://doi.org/10.1073/pnas.1711913115.

53. Nitzan, M., and Brenner, M.P. (2021). Revealing lineage-related signals in single-cell gene expression using random matrix theory. Proc. Natl. Acad. Sci. USA *118*, e1913931118. https://doi.org/10.1073/pnas.1913931118.

54. Pazos, F., Ranea, J.A.G., Juan, D., and Sternberg, M.J.E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. *352*, 1002–1015. https://doi.org/10.1016/j.jmb.2005.07.005.

55. Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., Shi, T., and Li, Y. (2005). Refined phylogenetic profiles method for predicting protein–protein interactions. Bioinformatics *21*, 3409–3415. https://doi.org/10.1093/bioinformatics/bti532.

56. Kann, M.G., Jothi, R., Cherukuri, P.F., and Przytycka, T.M. (2007). Predicting protein domain interactions from coevolution of conserved regions. Proteins *67*, 811–820. https://doi.org/10.1002/prot.21347.

57. Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc. Natl. Acad. Sci. USA *105*, 934–939. https://doi.org/10.1073/pnas.0709671105.

58. Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. *34*, 790–805. https://doi.org/10.1016/j.tig.2018.07.003.

59. Sul, J.H., Martin, L.S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet. *14*, e1007309. https://doi.org/10.1371/journal.pgen.1007309.

60. Hie, B.L., Yang, K.K., and Kim, P.S. (2022). Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. Cell Syst. *13*, 274–285.e6. https://doi.org/10.1016/j.cels.2022.01.003.

61. Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., and Kim, P.S. (2024). Efficient evolution of human antibodies from general protein language models. Nat. Biotechnol. *42*, 275–283. https://doi.org/10.1038/s41587-023-01763-2.

62. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

63. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

64. Edwards, U., Rogall, T., Blöcker, H., Emde, M., and Böttger, E.C. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res. *17*, 7843–7853. https://doi.org/10.1093/nar/17.19.7843.

65. Baker, G.C., Smith, J.J., and Cowan, D.A. (2003). Review and re-analysis of domain-specific 16S primers. J. Microbiol. Methods *55*, 541–555. https://doi.org/10.1016/j.mimet.2003.08.009.

66. Petti, C.A. (2007). Detection and Identification of Microorganisms by Gene Amplification and Sequencing. Clin. Infect. Dis. *44*, 1108–1114. https://doi.org/10.1086/512818.

67. Weisburg, W.G., Barns, S.M., Pelletier, D.A., and Lane, D.J. (1991). 16S ribosomal DNA amplification for phylogenetic study. J. Bacteriol. *173*, 697–703. https://doi.org/10.1128/jb.173.2.697-703.1991.

68. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

69. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

70. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

71. Haak, B.W., Littmann, E.R., Chaubard, J.-L., Pickard, A.J., Fontana, E., Adhi, F., Gyaltshen, Y., Ling, L., Morjaria, S.M., Peled, J.U., et al. (2018). Impact of gut colonization with butyrate-producing microbiota on respiratory viral infection following allo-HCT. Blood *131*, 2978–2986. https://doi.org/10.1182/blood-2018-01-828996.

72. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.

73. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321. https://doi.org/10.1093/sysbio/syq010.

74. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics *38*, 5315–5316. https://doi.org/10.1093/bioinformatics/btac672.

75. Lemoine, F., and Gascuel, O. (2021). Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. NAR Genom. Bioinform. *3*, lqab075. https://doi.org/10.1093/nargab/lqab075.

76. Rambaut, A., and Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. *13*, 235–238. https://doi.org/10.1093/bioinformatics/13.3.235.

77. Fraser, C., Hanage, W.P., and Spratt, B.G. (2007). Recombination and the nature of bacterial speciation. Science *315*, 476–480. https://doi.org/10.1126/science.1127573.

78. Van Rossum, T., Ferretti, P., Maistrenko, O.M., and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. *18*, 491–506. https://doi.org/10.1038/s41579-020-0368-1.

79. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

80. Lemoine, F., Domelevo Entfellner, J.B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature *556*, 452–456. https://doi.org/10.1038/s41586-018-0043-0.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
All sequencing data for commensal strain bank	This paper	NCBI (Accession: PRJNA737800) see Table S1 for individual accession numbers
Raw metabolic data for commensal strain bank	This paper	Metabolights (Study ID: MTBLS7771)
Experimental models: Organisms/strains		
E. rectale, Strain ID: MSK.9.13, NCBI_accession: JAAISA000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.9.15, NCBI_accession: JAAIRZ000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.13.48, NCBI_accession: JAAISJ000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.13.50, NCBI_accession: JAAISI000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.13.59, NCBI_accession: JAAISH000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.16.22, NCBI_accession: JAAIMQ000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.16.45, NCBI_accession: JAAIMP000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.13, NCBI_accession: JAAIMK000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.19, NCBI_accession: JAAIMJ000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.3, NCBI_accession: JAAIMG000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.42, NCBI_accession: JAAIME000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.57, NCBI_accession: JAAIMC000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.70, NCBI_accession: JAAILY000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.78, NCBI_accession: JAAILX000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.17.79, NCBI_accession: JAAILW000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.22.19, NCBI_accession: JAAISF000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.22.23, NCBI_accession: JAAISE000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.22.28, NCBI_accession: JAAISD000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.22.51, NCBI_accession: JAAISB000000000	This paper	Available upon request
E. rectale, Strain ID: MSK.22.92, NCBI_accession: JAJFBX000000000	This paper	Available upon request
Software and algorithms		
SpectralInference.jl	This paper	https://doi.org/10.5281/zenodo.13244626

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All collection of stool samples from healthy donors (leading to isolation of strains in Biobank) is covered under IRB 20-1384.

METHOD DETAILS

Genome sequencing of commensal strain bank

Fecal samples were obtained from 28 human donors that fell within the age range of 18 to 63 with a median age of 35. Donors were selected as those with no antibiotic use in the past year, no known history of diabetes, colitis, autoimmune disease, cancer, pneumonia, dysentery, or cellulitis at time of consent. Institutions that approved protocols of fecal sample collection were Memorial Sloan Kettering (MSK) and the University of Chicago under IRB 20-1384.

Fresh fecal samples were immediately reduced in an anaerobic chamber upon collection and diluted and cultured on various growth media. Agar media types vary, but include any of following: Columbia Blood Agar, Brain Heart Infusion +Yeast, Brain Heart Infusion + Mucin, Brain Heart Infusion + Yeast + Acetate or N-Acetylglucosamine, reinforced Clostridial Agar, Peptone Yeast Glucose, Yeast Casitone Fatty Acids, Defined media M5. Colonies were selected and grown to be sufficiently turbid, 20% glycerol/PBS stocks were created and stored in a -80°C freezer.

Colonies were selected for whole-genome sequencing based on pyro-sequencing of the 16S region which provides a rough estimate of genus level designation. For each donor, only colonies that had a sequence identity threshold of less than 99% from CD-Hit (v. 4.8.1) were selected for whole-genome sequencing.[62] Bacterial genomic DNA was extracted using QIAamp DNA Mini Kit (QIAGEN) according to manufacturer's manual. The purified DNA was quantified using a Qubit 2.0 fluorometer. 1000ng of each sample was prepared for sequencing using the QIAseq FX DNA Library Kit (QIAGEN). The protocol was carried out for a targeted fragment size of 550bp. Sequencing was performed on the MiSeq or NextSeq platform (Illumina) with a paired-end (PE) kit in pools designed to provide 1-3 million PE reads per sample with read length of 250 or 150 bp. Adapters were trimmed off with Trimmomatic with following parameters: the leading and trailing 3 bp of the sequences were trimmed off, quality was controlled by a sliding window of 4, with an average quality score of 15 (default parameters of Trimmomatic). Moreover, any read that was less than 50 bp long after trimming and quality control were discarded. The remaining high-quality reads were assembled into contigs using SPAdes (v3.14.0).[63] The primers associated with 16S and whole genome sequences are as follows:

Primer pair	Forward primer (5'–>3')	Reverse primer (5'–>3')	Trials
8F-1492R[64,65]	AGA GTT TGA TCC TGG CTC AG	GGT TAC CTT GTT ACG ACTT	16S rRNA full gene length
533F-907R[66,67]	GTG CCA GCA GCC GCG GTA A	CCG TCA ATT CMT TTR AGT TT	Sanger Sequencing, V4-V5

Taxonomic classification of the assembled contigs was performed with the following methods: (a) Kraken2 (v2.1.1); (b) full/partial length 16S rRNA gene from each isolated colony's assembled contigs is extracted and input into BLASTn (v2.10.1+) to query against NCBI's RNA RefSeq database.[68,69] Top five hits for each query are manually curated to determine an isolate's identity, with identity and coverage cutoff both at 95%; (c) GTDB-Tk (v1.5.1).[70] Final taxonomy is determined by the consensus of the three methods. Any colony that did not match initial pyro-sequencing taxonomy or lacked consensus are excluded from the commensal strain bank.

Metabolic profiling of strains

Strains were grown in Brain Heart Infusion media supplemented by cysteine (BHIS) until sufficiently turbid and then spun down. Supernatant samples were frozen at -80°C prior to extraction. Samples were thawed and 4 volumes of extraction solvent (100% methanol spiked with internal standards: D_6-succinate (1 mM), D_5-phenol (0.025 mM)) was added to the liquid sample (1 volume) in a microcentrifuge tube. The raw peak area of the internal standards were averaged for peak normalization. Tubes were then centrifuged at -10 °C, 20,000 x *g* for 15 min and supernatant was used for subsequent metabolomic analysis. Compounds were derivatized with pentafluorobenzyl bromide (PFBBr) as described by Haak et al. with the following modifications.[71] The metabolite extract (100 μL) was added to 100 mM borate buffer (100 μL, pH 10), 100 mM pentafluorobenzyl bromide in acetonitrile (400 μL), and *n*-hexane (400 μL) in a capped mass spectrometry autosampler vial. Samples were heated in a thermomixer C (Eppendorf) to 65 °C for 1 hour while shaking at 1300 rpm. After cooling to room temperature, samples were centrifuged at 4 °C, 2000 x g for 5 min, allowing phase separation. The hexanes phase (100 μL) (top layer) was transferred to an autosampler vial containing a glass insert and the vial was sealed. Another 100 μL of the hexanes phase was diluted with 900 μL of *n*-hexane in an autosampler vial. Concentrated and dilute samples were analyzed using a GC-MS (Agilent 7890A GC system, Agilent 5975C MS detector) operating in negative chemical ionization mode, using a HP-5MSUI column (30 m x 0.25 mm, 0.25 μm; Agilent Technologies 19091S-433UI), methane as the reagent gas (99.999% pure) and 1 μL split injection (1:10 split ratio). Oven ramp parameters: 1 min hold at 60 °C, 25 °C per min up to 300 °C with a 2.5 min hold at 300 °C. Inlet temperature was 280 °C and transfer line was 310 °C. Data analysis was performed using MassHunter Quantitative Analysis software (version B.10, Agilent Technologies) and the 50 targeted compounds—spanning SCFA, BCFA, amino acid, aromatic, hydroxylated fatty acid, organic acid, indole, and additional subclasses—were identified by comparison to authentic standard *m/z*, retention time and fragmentation pattern. Normalized peak areas were calculated by dividing raw peak areas of targeted analytes by averaged raw peak areas of internal standards. The compounds chosen within the PFBBR panel represent mechanisms known to be important in health and disease and were compiled from well-known mechanisms in literature and human, murine, and *in vitro* datasets collected within the Duchossois Family Institute (DFI).

Motility assay for E. rectale isolates

On day 1, BHIS media (250mL dH2O, 9.25g BHI Media, 2.5mL cysteine solution (1g cysteine in 10mL dH2O)) was made and aliquoted into 50 mL conical tubes. Tubes were cycled into anaerobic chamber (Coy) 24 hours prior to the experiment. Caps on tubes were left loose to allow for equilibration to the anaerobic environment and to release excess oxygen that may impact strain growth. On day 2, a serological pipet was used to aliquot 5 mL BHIS into 20 mL conical tubes. Glycerol stocks of *E. rectale* strains were cycled into the anerobic chamber; media was inoculated with strains in triplicate and placed in a 37° incubator in the anaerobic chamber. OD_{600} was measured every 12 hours for 48 hours for each tube of inoculated media by sampling from the top of the culture (taking 100μL from the top 1mL of the 5mL cultures). During this time, each tube was also observed for pellet formation. After 48 hours of incubation, cultures were briefly vortexed to disseminate any pellet formed at the bottom of the tube. Samples were then collected from the top of each culture (100μL from the top 1mL) and measured for their OD_{600} every thirty minutes for 180 minutes after vortex.

QUANTIFICATION AND STATISTICAL ANALYSIS

Phylogenetic trees of strain bank

The 16S sequence was isolated from each strain in the commensal strain bank. All 16S sequences were aligned with Mafft (v7.520) creating a multiple sequence alignment of 1521 features and 335 unique sequences.[72] This alignment was then input to phyml (v3.3.20200621) with these command options: *"./phyml -dnt -mHKY85 -fe -o tlr –search SPR –r_seed 123456 –rand_start –n_rand_starts 3 –no_memory_check –bootstrap -4 -i BB669_16S.phy".*[73] Redundant sequences were placed into the final tree by PhyML at distance zero from their identical representative in the tree.

For Bac120, the fastafiles for each isolate was input to the gtdbtk (v2.3.0) 'identify and align' pipeline to create a multiple sequence alignment with the Bac120 feature set comprising 5,035 features, with 311 unique sequences.[74] This alignment was input to PhyML (v3.3.20200621) with these command parameters: *"./phyml -daa -mLG -fe -o tlr –search SPR –r_seed 123456 –rand_start –n_rand_starts 3 –no_memory_check –bootstrap -4 -i BB669_Bac120.phy"*.[73] Redundant sequences were placed into the final tree by PhyML at distance zero from their identical representative in the tree.

Spectral distance and spectral groups

Spectral groups and spectral distance are based on Singular Value Decomposition (SVD), which is a matrix factorization method that is a generalization of Principal Components Analysis (PCA). In the main text, we use the term 'principal components'; we note that principal components are also called 'spectral components', 'modes of variation', or 'eigenmodes' in the literature. These terms all describe the same mathematical concept.

In general, SVD factorizes a real matrix **M** into three matrices according to the following equation:

$$M = U\Sigma V^T \qquad \text{(Equation 2)}$$

In Equation 2, **U** is termed the left-singular vector (LSV) matrix, **Σ** is a diagonal matrix of singular values, and **V** is termed the right-singular vector matrix. If **M** is a matrix of *n* systems (rows) described by *m* features (columns) where $n < m$; **U** is an $n \times n$ matrix where rows are systems, columns are LSVs, and each entry is the contribution of a given system to an LSV; **Σ** is an $n \times m$ matrix where the k^{th} diagonal entry is the k^{th} singular value and all off-diagonal entries are 0; and **V** is an $m \times m$ matrix where rows are features, columns are RSVs, and each entry is the contribution of a given feature to an RSV. $\mathbf{V^T}$ in Equation 2 is the transpose of **V**. A 'spectral component' is the axis specified by the k^{th} singular value and is the same as a 'principal component' from PCA, or 'eigenmode', or 'mode of variation'. The relationship between SVD and PCA is that PCA is performed only on either the rows or the columns of **M**. Therefore, matrices **U** and **Σ** can be multiplied together to form **P** which are exactly the principal components of matrix **M**.

$$P = U\Sigma \qquad \text{(Equation 3)}$$

As an example, the diversification trajectory shown in Figure 1A is a representative of the toy models we used to conduct our analysis. In Figure 1A, each taxon in the alignment is defined by a 'genotype' comprised of fourteen features that are either a '1' or a '0'; each taxon is created from a series of three sequential diversification events. Collectively, the alignment of taxa represents extant diversity. The ancestral root is defined as a genotype of all '1'. The first layer of diversification from the ancestral root is defined by two separate mutations in positions 1 and 2. The second layer of diversification mutates positions 3 through 6 to create sub-populations. The third layer of diversification mutates positions 7 through 14 to create the extant diversity of taxa—eight taxa in total with diverse genotypes.

PCA on the alignment of taxa yielded eight spectral components (Figure 1B). The extent to which each taxon contributes, or 'projects' onto each spectral component is shown in Figure 1C. When visualizing the contribution of each taxon onto each spectral component, we observe that taxa arising from a common broad layer of diversification contribute similarly to the first two spectral components while those arising from common finer layers of diversification (the second and third diversifications) contribute similarly to deeper spectral components.

We translated our finding into a mathematical entity by computing the 'spectral distance' between two taxa. The spectral distance between two taxa, ***i*** and ***j***, on spectral component ***k*** is

$$SD_{ij}^{k} = \left| P_i^k - P_j^k \right| \qquad \text{(Equation 4)}$$

where $\boldsymbol{P_i^k}$ is the projection of taxa ***i*** onto spectral component *k* and $\boldsymbol{P_j^k}$ is the contribution of taxa ***j*** onto spectral component *k*. We show an example of a pattern of spectral distances in Figure 1D where taxon 'a' is the reference. See that taxon 'a' and all other taxa share the same projection onto the first spectral component. As such, the spectral distance between 'a' and any other taxa is zero at spectral component 1. However, at spectral component 2, 'a' continues to share the same projection as taxa 'b', 'c', and 'd', but taxa 'e', 'f', 'g', and 'h' have a different projection onto spectral component 2. Therefore, the spectral distance between 'a' and 'e', 'f', 'g', and 'h' is non-zero at spectral component 2 but still zero for 'b', 'c', and 'd'.

We next defined the 'cumulative spectral distance' between two taxa from the first to k^{th} spectral component as

$$SD_{ij}^{1:k} = \left| P_i^k - P_j^k \right| + \sum_{r=1}^{k-1} \left| P_i^r - P_j^r \right| \qquad \text{(Equation 5)}$$

where *r* denotes the index of each spectral component 'shallower' than spectral component *k* (i.e., each spectral component with an associated singular value greater than the k^{th} singular value). The cumulative spectral distance pattern of all pairs of taxa with taxa 'a' as one member of the pair is shown in Figure 1D. Computing the cumulative spectral distance across all spectral components for all pairs of taxa illustrated a distinct tree-like pattern of partitioning between taxa.

Our key finding from Figure 1D, that groups of sequential spectral components collectively described different layers of precedent diversifications, motivated 'renormalizing' the spectral components into groups of spectral components harboring the same percent data-variance (Figure 1E, left panel). Spectral component groups were thus defined based on the natural $\log_{10}$ difference between subsequent singular values. The rationale behind this choice was to group together spectral components that are relatively similar in their measure of percent-variance explained. To reduce effects at very small singular values, a pseudo-count of 1 is added to each singular value. The difference between each subsequent singular value is expressed as

$$\Delta_k = \ln(\Sigma_{k-1,k-1} + 1) - \ln(\Sigma_{k,k} + 1) \qquad \text{(Equation 6)}$$

Then, the k^{th} component was chosen to start a new spectral group only if the difference between the $(k-1)^{th}$ and k^{th} component was greater than a manually chosen threshold θ,

$$K = \{k \mid \Delta_k > \theta\} \qquad \text{(Equation 7)}$$

For our *in-silico* toy models, the threshold θ was 0 and therefore any drop in explained variance was defined as a new group of spectral components. For real biological data described in the main text, the threshold θ was chosen as 1.5 times the third quantile of these natural log differences ($\theta = 1.5 \times Q3(\Delta)$) as an approximation for selecting the only the largest drops in explained variance. The spectral distance computed across spectral component groups was defined as

$$SD_{ij}^{G} = \sum_{g \in G} \left| P_i^g - P_j^g \right|_{l2} \qquad \text{(Equation 8)}$$

where G is the total set of spectral component groups, g is a specific spectral component group within G, and $|\cdot|_{l2}$ denotes the l_2 norm also known as the Euclidean distance.

Measuring the accuracy of Spectral Trees

We sought to assay the robustness of Spectral Trees to (i) size of the alignment and (ii) number of features describing each system in the alignment. To perform this test, we used GoTree v0.4 (https://github.com/evolbioinfo/gotree) to create 7 separate reference 'ground truth' trees of taxa comprising either 16, 32, 64, 128, 256, 512, or 1,024 leaves.[75] Each tree was then input into SeqGen v 1.3 (https://github.com/rambaut/Seq-Gen) which produced multiple sequence alignments (MSAs) where rows were leaves and columns were features describing the leaves.[76] SeqGen uses a Markov process considering the branching pattern of the tree to create a vector of features for each leaf. Elements of the vector are the characters 'A' and 'T', and the Markov process uses uniform probabilities to flip between these characters at each branch point in the tree. For each of the 7 trees, we generated 7 MSAs with SeqGen, where each MSA contained either 16, 32, 64, 128, 256, 512, and 1024 features. Thus, our analysis spanned 49 total MSAs.

We compared the topology of Spectral Trees against ground-truth defined by GoTree using an F-score—the harmonic mean of precision and recall. Precision between two trees is defined as the proportion of predicted branches in the Spectral Tree that are also in the 'true' tree. Recall is defined as the proportion of branches in the 'true' tree that are also in the Spectral Tree. F-score ranges between 0 to 1, where 1 indicates complete identity between the two trees and 0 indicates no commonality between the two trees.

Our results are shown in Figure S3. We found that for the majority of the parameter space, the F-statistic was near 1. In the limit that the number of features was less than the number of taxa, the F-statistic was uniformly near 0. This distinction in F-statistic based on the parameter space arises from the scenario where the number of features is the limiting descriptor relative to the number of taxa in the alignment. The physical interpretation of this regime is that the number of features describing each system is substantially limited compared to the diversity of systems available for sampling. In this case, the information content of the set of features is 'overwritten' by the diversity of taxa, thereby erasing patterns of covariation originating from phylogenetic histories. A biological process that is consistent with this regime is if the recombination rate is extremely high relative to speciation events—a scenario that has been put forth as a plausible scenario for bacterial phylogenomic trends.[31,77,78]

Mutual Information (MI) calculation

We sought to elucidate where information regarding the different generations of diversifications lay across the set of spectral components. We conducted this analysis by first defining sequential windows of spectral components across all nine spectral components (components 1 to 3, 2 to 4, …, 7 to 9). For each spectral window, we isolated the corresponding LSVs from the ***U*** matrix defined by Equation 2. This results in several sub-matrices defined by taxa on the rows, LSVs on the columns, and each entry being the contribution of each taxon onto each LSV. For each sub-matrix constructed from ***U***, we computed the Spearman correlation between all pairs of taxa across the set of LSVs defined in the sub-matrix ('spectral correlations'). As a concrete example, for the first spectral window comprising spectral components 1 to 3, the ***U*** submatrix is defined as taxa (rows) and the first three columns (LSV1 to LSV3) of the ***U*** matrix. Then, to compute the spectral correlations between all pairs of taxa within LSVs 1 to 3, we computed the Spearman correlations between all pairs of rows in the sub-matrix. The result is a taxon-by-taxon spectral correlation matrix where each entry is the Spearman correlation measured between two taxa across spectral components 1 to 3. Defining all taxon-by-taxon spectral correlation matrices across all spectral windows creates a three-dimensional tensor, **R**, defined by taxa (rows), taxa (columns), and spectral windows (z-axis) where each entry in the tensor is the spectral correlation between two taxa within a spectral window. Separately, we created a second tensor, **G**, where rows are defined as taxa, columns are defined as taxa, the z-axis is each generation ('F0', 'F1', 'F2', or 'F3'), and entries in the tensor are a '1' if two taxa are grouped within the same cluster at a given generation or ‘0’ if two taxa are not grouped within the same cluster at a given generation. We then computed the mutual information (MI) between each face of the R tensor with each face of the G tensor. This computation interrogated the information shared between spectral correlations between taxa and shared generational history. The MI was calculated as

$$MI(r|G) = H(r) - \left(\frac{n_0}{N}H(r_0) + \frac{n_1}{N}H(r_1)\right) \qquad \text{(Equation 9)}$$

where *H(r)* is a measure of entropy and is defined as

$$H(x) = \log_2(\Delta_{bw}) - \sum_b p(x_b)\log_2(p(x_b)) \qquad \text{(Equation 10)}$$

$p(x_b)$ is the proportion of pairs that fall into a particular bin within a distribution of x_b values; we use a bin-width of 0.01 Δ_{bw} to construct 200 bins across the distribution of correlation values ranging from -1 to 1; r_1 is the distribution of spectral correlations across taxonomic pairs that are descendants of the same ancestor; n_1 is the number of pairs in r_1; r_0 is the distribution of spectral correlations across taxonomic pairs that are not descendants of the same ancestor; n_0 is the count of those pairs within each bin; *N* is the total number of pairs. The meaning of this calculation is a measure of the extent to which knowing the distribution of spectral correlations within a spectral window between two taxa indicates the shared ancestral history of two taxa.

Recreating the alignment

We leveraged a central property of Singular Value Decomposition (SVD) and PCA—their linearity—to isolate the statistical information in distinct principal components. We can rewrite Equation 2 as

$$M = \sum_k \sigma_k u_k v_k^t \qquad \text{(Equation 11)}$$

where σ_k is the k^{th} singular value and u_k and $v^t{}_k$ are the k^{th} left and right singular vectors. Each product that is being summed in Equation 11 is a rank 1 matrix because it produces a matrix from the individual vectors of $u_k v^t{}_k$ that is scalar multiplied by the number σ_k. Using Equation 11, we recreated the original alignment shown in Figure 2A but only considering the information contained in principal components 5 to 8. This process is shown in Figure 2F. Focusing on position 5 again, we found that the value of position 5 was adjusted in the recreated alignment reflecting the separate, nested contexts of diversification (Figure 2G). Thus, by considering information contained across all principal components, the Spectral Tree accurately resolved both broad and context-dependent, finer patterns of diversification.

Creating a Spectral Tree across UniProt

Construction of the full alignment of 7,047 UniProt reference proteomes annotated by 10,177 Orthologous Gene Groups was previously described in Zaydman et al.[30] Then using this alignment, inferred trees of taxonomic relatedness (‘Spectral Trees’ in the main text) are generated using four steps.

Step 1: A reference pairwise spectral distance matrix **SD** is created from Equation 8 for all pairs of taxa comprising a matrix **M**.

Step 2: A reference tree $\mathbf{ST_{ref}}$ is generated via hierarchical clustering using the NeighborJoining method of phylogenetic tree building[79]

Step 3: A set of 100 bootstrap trees is generated using steps 1-2. For each bootstrap, we replace the original matrix **M** with a bootstrap matrix $\mathbf{M_{boot}}$ by sampling features (columns) with replacement to maintain the original dimensions of **M**. This procedure first generates a pairwise distance matrix $\mathbf{SD_{boot}}$ from the matrix $\mathbf{M_{boot}}$, and then generates a tree $\mathbf{ST_{boot}}$ using the NeighborJoining algorithm.

Step 4: The reference tree $\mathbf{ST_{ref}}$ and the bootstrap trees are then compared with transfer bootstrap expectation (TBE) as described in Lemoine et al.[80] TBE ranges from 0 to 1 where 0 indicates no similar branches in any bootstrap tree, and 1 indicates the exact branch was found across all bootstrap trees.

The result of implementing these four steps generates a Spectral Tree where each branch of the tree has an associated measure of support as defined by TBE. The Spectral Tree associated with the alignment in Figure 3A is shown in Figures 3A and 3C.

MI between Spectral Tree and phylogeny

We first created 100 ‘cuts’ of the tree where each cut is equally spaced across the depth of the tree. The first cut is defined at the root of the tree forming a single cluster comprising all taxa; the last cut is defined at the terminal branches of the tree forming as many clusters as there are taxa. For each cut, we form two membership vectors **C** and **T** where each element in the vector represents a pair of taxa in the tree. **C** is a ‘1’ if two taxa belong to the same tree cluster and a ‘0’ if not. **T** is a ‘1’ if two taxa share a property (i.e. belong to the same NCBI taxonomic designation or come from the same donor) and a ‘0’ if not. **T** is constructed for (i) all taxonomic designations spanning ‘Phylum’ to ‘Species’ in NCBI, (ii) all taxonomic designations spanning ‘Phylum’ to ‘Species’ in GTDB, and (iii) identity of donor. We then calculate the mutual information (MI) between **C** and **T** by the following equation:

$$MI(\boldsymbol{C},\boldsymbol{T}) = H(\boldsymbol{C}) + H(\boldsymbol{T}) - H\left([\boldsymbol{C},\boldsymbol{T}]^t\right) \qquad \text{(Equation 12)}$$

where $H(x) = -\sum_{b \in 0,1} p(x_b)\log_2(p(x_b))$ represents the Shannon entropy; $p(x_b)$ is the proportion of x that is equal to either '0' or '1' respectively in the distributions defined by either **C** or **T**. $H([\boldsymbol{C}, \boldsymbol{T}]^t)$ is the joint entropy of **C** and **T**. The 'Cumulative MI density' plotted in Figures 3B and 4D is defined by adding the MI for each subsequent deeper cut of the tree and dividing by the total sum of MI across all cuts. NCBI phylogenetic strings were mapped to NCBI taxonomy IDs following methods described in Zaydman et al.[30]

To measure uncertainty in MI, we bootstrapped the MI calculations. For a given MI between **C** and **T**, pairs of taxa (matched elements of **C** and **T**) are sampled with replacement to the same total number of pairs. As an example, for 10 choose 2 pairs (n=45), 45 pairs are sampled with replacement. This bootstrapping is performed 50 times and the cumulative mean ± 2 standard deviations of MI is plotted as ribbons.

Projecting CSB into the Spectral Tree

Genome sequences of all strains from the commensal strain bank were put through EggNog mapper (emapper 5.0) and their proteomes were annotated for their OGG content across the same set of OGGs defining the UniProt reference proteome database (n = 10,177).[34,35] Any OGG measured in UniProt but not in the commensal strain bank was imputed as a '0' count. From Equation 2, we calculated the principal components defining covariation amongst the UniProt reference proteomes, $P^{UniProt}$, as

$$P^{UniProt} = D^{OGG} V^{UniProt} \qquad \text{(Equation 13)}$$

where D^{OGG} is the 7,047 UniProt reference bacterial proteomes annotated by their 10,177 OGGs. We next defined B^{OGG} as the matrix of commensal strain bank strains annotated by their OGGs. Therefore from Equation 3,

$$P^{CSB} = B^{OGG} V^{UniProt} \qquad \text{(Equation 14)}$$

where P^{CSB} is a matrix of commensal strains (rows) by 7,047 principal components (columns) that collectively define the structure of bacterial co-evolution in the UniProt database; each entry is the contribution of each commensal strain bank strain onto each principal component.

SLE LASSO model training and validation

To establish a training and validation set, we selected all strains belonging to the 11 species with 20 or greater biological replicates (n=356), and then further subset to 75% of those strains maintaining relative proportions of species groups (n = 267) with the remaining 25% (n = 89) used as a validation set.

To train the LASSO model, we first generated a Spectral Tree from the training set and created an associated SLE matrix ($\mathbf{SLE_{train}}$, 267 rows by 266 columns) per the diagram in Figure 6A. Each strain was labeled by the fold-change (log2FC) of a specific metabolite. We next estimated the linear coefficients relating SLEs with relative change in metabolite concentration across varying degrees of regularization by

$$\widehat{w} = \underset{w}{\operatorname{argmin}}\left(\left|\boldsymbol{SLE_{train}}\, w - y\right|_2 + \lambda |w|_1 \right) \qquad \text{(Equation 15)}$$

where $\widehat{w}$ is the estimated coefficients, y is the log2FC of a metabolite, and λ is the regularization parameter swept from 10^0 to 10^{-3}.

We made predictions for the validation set in two steps. First, we found the nearest neighbor in the training set for each validation strain using spectral distance. Second, we input the SLE of the nearest neighbors into the SLE-LASSO model trained via Equation 15. Collectively, the out-of-fold predictions for one statistical resampling are described by

$$\widehat{y} = \boldsymbol{SLE_{train}}\, \widehat{w} \qquad \text{(Equation 16)}$$

where $\widehat{y}$ represents the out-of-fold predictions for a single fold; SLE_{train} contains the nearest neighbors in the training set to each test set strain; and $\widehat{w}$ are the SLE-LASSO coefficients learned from the full training set. This setup for making out-of-sample predictions is similar to a K-nearest-neighbors model in that we use the features contained in our training set to make predictions. However, by incorporating learnable weights for different sections of the Spectral Tree we tune the number of neighbors the prediction is averaged across dependent on the phylogenetic context of each strain.

We repeat creating a training set, creating an associated Spectral Tree, creating an associated $\boldsymbol{SLE_{train}}$, and making out-of-fold prediction across 20 repartitions comprising 4-folds per data partition across 5 re-partitions. This validation procedure guarantees 5 out-of-fold predictions per taxa. We measure the performance of the model using the out-of-fold adjusted R^2 for each repartitioning.

Separability of donor metabolic variation

Each of the 7,040 models trained in STAR Methods: SLE LASSO model training and validation—one model for each metabolite (n=32), species (n=11), and repartition of train-test sets (n=20)—was tuned with regularization parameter λ to maximize predictive capacity (adjusted R^2) and sparsity (maximum λ for given maximum predictive capacity) for each metabolite-species pair. In Figure 7G, the maximized out-of-fold adjusted R^2 per species was plotted on the y-axis. For each of the 7,040 out-of-fold test sets, we calculated a 'separability index'

$$\log_2\left(\frac{\text{inter-donor variation}}{\text{intra-donor variation}}\right) \qquad \text{(Equation 17)}$$

where the numerator, 'inter-donor variation'. is the standard deviation of the mean metabolite concentration for strains belonging to each donor, and the denominator, 'intra-donor variation', is the mean of the standard deviations of metabolite concentration for strains belonging to each donor. The ratio between inter-donor and intra-donor variation in metabolite concentrations defines how separable the measured metabolite concentrations are by knowing which donor a strain of a given species is collected from. To control for cases where there is very low or no measured metabolite for a given metabolite-species-resample test set a regularization constant of $\frac{1}{2^7}$ is added to both the numerator and denominator before taking the $\log_2$. The final separability index is defined as the $\log_2$ value of the ratio. Values greater than 0 indicate the metabolite concentrations are separable by donor; values less than zero indicate that the metabolic variability of strains within a donor is inseparable from the distributions of other donors.

Significance of enrichment or depletion for models relative to adjusted r^2 value of models was performed with a Fisher's exact test with Bonferroni correction on a contingency table that tallies the number of models with positive versus negative predictive capacity and positive versus negative separability index. The tallied values are compared against the expected counts under the null of no association between predictive capacity and separability index. P-values are reported in Figure 7G.