AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Computational Biology | Research Article

Augmenting bacterial similarity measures using a graph-based genome representation

Vivek Ramanan,[1,2] **Indra Neil Sarkar**[1,2,3]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT Relationships between bacterial taxa are traditionally defined using 16S rRNA nucleotide similarity or average nucleotide identity. Improvements in sequencing technology provide additional pairwise information on genome sequences, which may provide valuable information on genomic relationships. Mapping orthologous gene locations between genome pairs, known as synteny, is typically implemented in the discovery of new species and has not been systematically applied to bacterial genomes. Using a data set of 378 bacterial genomes, we developed and tested a new measure of synteny similarity between a pair of genomes, which was scaled onto 16S rRNA distance using covariance matrices. Based on the input gene functions used (i.e., core, antibiotic resistance, and virulence), we observed varying topological arrangements of bacterial relationship networks by applying (i) complete linkage hierarchical clustering and (ii) K-nearest neighbor graph structures to synteny-scaled 16S data. Our metric improved clustering quality comparatively to state-of-the-art average nucleotide identity metrics while preserving clustering assignments for the highest similarity relationships. Our findings indicate that syntenic relationships provide more granular and interpretable relationships for within-genera taxa compared to pairwise similarity measures, particularly in functional contexts.

IMPORTANCE Given the prevalence and necessity of the 16S rRNA measure in bacterial identification and analysis, this additional analysis adds a functional and synteny-based layer to the identification of relatives and clustering of bacteria genomes. It is also of computational interest to model the bacterial genome as a graph structure, which presents new avenues of genomic analysis for bacteria and their closely related strains and species.

KEYWORDS synteny, genome analysis, microbiome

16S ribosomal RNA regions are used to identify bacteria and form the foundation for phylogenetic relationships between bacterial groups (1, 2). 16S rRNA analyses use variable regions of the 16S region to identify groups of similar sequences (3). The level of identification (e.g., strain, species, and genus) depends on sequencing power. However, improvements in technology expand the sequencing potential beyond the variable region of the 16S gene to the entire 16S region, as well as to entire genomes (4). Studies have shown that 16S-based analysis is not infallible and does not always corroborate other forms of phylogenetics or taxonomy (5, 6). 16S analyses rely on reference databases and heuristic clustering into "operational taxonomic units", which can remove individual genomic sequences in favor of a consensus sequence (7). While 16S regions continue to be the leading form of identification in bacteria, there are also numerous pairwise data that are often analyzed post-16S identification or separate from 16S analysis. There is, therefore, potential to combine or analyze 16S alongside other data.

Editor Sergio Baranzini, University of California, San Francisco, California, USA

Address correspondence to Indra Neil Sarkar, neil_sarkar@brown.edu.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 16 April 2024
Accepted 5 June 2024
Published 28 June 2024

Matrix transformation is a mathematical approach that combines a pairwise matrix with another equally sized matrix. This approach transforms original data into a matrix that contains the variation of both input matrices enabling the combination to be analyzed as a single matrix. In the context of 16S, full-genome information offers the potential for additional types of pairwise data. One example of the phenomenon of sequencing improvement is full-genome alignments, possible with tools such as MASH or FastANI (8, 9). Here, entire genomes can be aligned in computationally efficient ways to gather similarity or distance scores between a pair of genomes. A limitation of this strategy is that distance scores are only provided under a certain threshold, which affects pairs that are not closely related. Contemporary metrics such as SKANI also include the usage of orthologous segments in average nucleotide identity (ANI) calculation, improving clustering quality of within-species phylogeny but is limited to calculating 82% ANI and higher (10).

Synteny is a comparative genomics approach that aligns shared segments of DNA between a pair of genomes, highlighting differences in a shared segment location (11). The syntenic approach has been used as a visual representation for rearrangements, discovery of new shared segments, DNA order changes due to evolution, and genomic dynamics of subspecies (12–14). A range of tools exist that provide synteny block construction (e.g., Sibelia) and visualization (e.g., Synteny Portal) (15, 16). Synteny is used in targeted analyses and, to date, has not been employed in large-scale analyses across a representative set of bacteria. Given synteny can provide a representation of similarity between two related genomes, we propose a twofold approach for analyzing synteny: (i) characterize synteny as a representation of similarity between two related genomes and (ii) use syntenic data as an augmentation to 16S data using matrix transformation. Our synteny measure, which is based on a geometric graph structure, uses orthologous genes as the connection points between a given pair of genomes. We test the impact of this measure by transforming 16S rRNA data to demonstrate changes in clustering and graphical results, which can be used to understand new relationships between bacteria based on different data contexts. We compare this with state-of-the-art techniques with ANI calculation to test the difference in both clustering quality and clustering results.

MATERIALS AND METHODS

Data acquisition from GenBank and ortholog construction

Bacterial genomes from GenBank were downloaded using ncbi-genome-download (17). For each bacterial species, one strain was chosen at random, after which only genomes with 16S genes were chosen. CheckM was used to evaluate genome completeness and contamination, of which genomes with completeness >90% and contamination <5% were used for the study (18). This resulted in 378 bacterial genomes in total for analysis. Genomes were organized in both GenBank Flat File and FASTA formats for further analysis. Taxonomic data on each species were gathered using NCBI Taxonomy.

Core gene ortholog construction

Core genes were identified using the UBCG2 data set (19). Core gene names and functions were taken from UBCG2 and compared with annotated genes and gene functions in GenBank flat files to identify core genes from the database. Identified genes were BLASTed against the entire gene database to identify orthologs, of which orthologs with greater than 95% nucleotide identity and a base pair length between 500 and 2,500 base pairs were kept.

Comparative distance score calculation: 16S, MASH, and ANI

16S rRNA genes were identified from GenBank annotated flat files, which have been computationally predicted using protein homology. These genes were compared to the SILVA ribosomal RNA gene database project to confirm 16S identity (20). For each pair

of genomes, the MASH distance between each set of 16S genes per gene was calculated (8). Whole genome MASH distances were also calculated using FASTA versions of sample genomes and ANI values calculated using SKANI on FASTA files. MASH distances were organized into a pairwise distance matrix, where the column and row indicated the species pair. The MASH distance was also validated against known taxonomic distance based on species name.

Synteny graph structure and measure

The construction of the pair synteny graph requires (i) a relatively ordered set of genes for each of the genomes and (ii) a set of at least two orthologous gene pairs between the two genomes. First, two linear graphs are created for each genome, after which edges are drawn between the orthologous gene pairs. Next, the cosine similarity of every combination pair of orthologs is calculated, using the relative order position of the gene in genome A and the gene in genome B (Fig. 1). The average of the array of cosine similarities is calculated as the synteny similarity for that arrangement. To perform the other rearrangements to simulate the circularity of genomes, a single pair is chosen as the pivot and arranged for each of the pair genes to be the top gene of the linear genome graph. The other genes are respectively ordered underneath the pivot gene of the pivot pair. The cosine similarities of every pair of orthologs are calculated, and the average cosine similarity is produced. To find the final synteny similarity, the median synteny similarity across all rearrangements is used. Median synteny similarity for each pair of genomes was organized into a pairwise similarity matrix, of which the distance matrix was calculated by subtracting every matrix value from 1. Only core gene

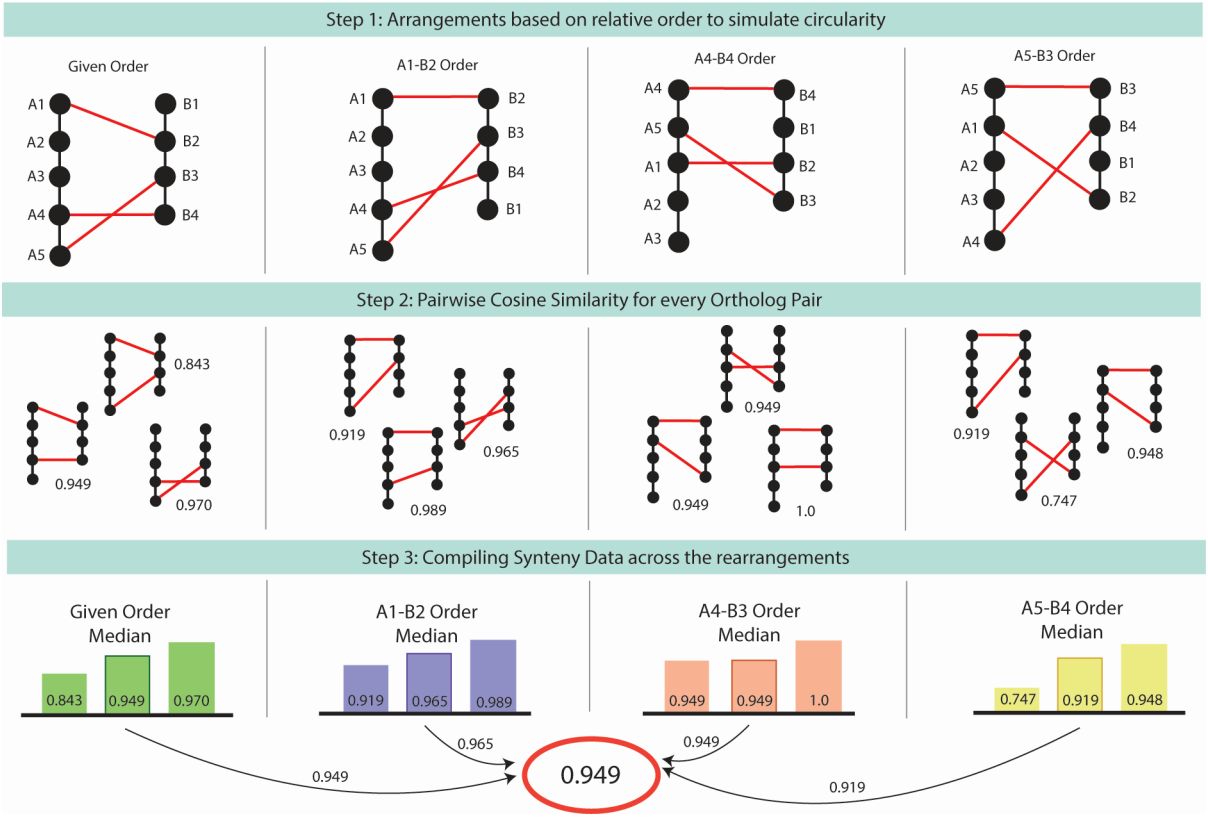

FIG 1 Visualization of the synteny similarity process. In step 1, multiple arrangements are created based on the ortholog set given, where the first order is the given order from the GenBank flat files and each following order is based on a chosen ortholog as the pivot. In step 2, the cosine similarity for each pair of orthologs is calculated, where values closer to 1 indicate similar positions of the orthologs and values closer to 0 indicate lower similarity. Finally, in step 3, the values are compiled per arrangement to find the median; then, the median among all the arrangements is chosen. Median values were used to reduce skew based on a wide range of cosine similarity values.

orthologs were used in the first iteration of the synteny similarity matrix, after which other functional gene group orthologs were used.

Synteny coverage metric

Synteny coverage was calculated as an asymmetric pairwise measure, equaling the summed length of the shared nucleotide blocks of a pair divided by the total genome length of the chosen genome (11). Therefore, this measure is different for each of the genomes used in the pairwise measure.

Random forest prediction

Random forest models were trained on a combination of 16S, whole genome distance, or syntenic distance (with all combinatoric possibilities) to predict taxonomic distance. Taxonomic distance was formed as a binary variable for each Linnean taxonomic level (i.e., kingdom, phylum, class, order, family, and genus), and an individual model was trained for each level. Accuracy and Receiver Operating Characteristic (ROC) curves were calculated for every model possibility.

Augmenting 16S data with syntenic similarity

To perform augmentation of the 16S MASH distance matrix, we used the covariance matrix of the synteny distance matrix. The dot product of the 16S MASH distance matrix and the covariance matrix of the synteny distance matrix resulted in the augmented 16S synteny distance matrix, which was normalized to range between 0 and 1. 0 indicated complete similarity based on 16S and synteny data, while 1 indicated no similarity. The completed matrix is symmetric, in which the diagonals indicate complete similarity for the synteny of the same genome.

Hierarchical clustering

Complete linkage hierarchical clustering was performed on the synteny-scaled 16S distance matrix, the original 16S distance matrix, and the ANI distance matrix. Dendrograms were visualized using unrooted trees in ggTree and labeled using the bacterial taxonomic phyla (21). The cluster cutoff number was varied to analyze metrics over a single experimental variable. To compare the results of the clustering groups, the Rand score and silhouette scores were all calculated. Rand scores form a similarity score for two clusterings when the matches between clusters are not known, by accounting for all combinations of cluster pairs between the two. Rand scores range from 0 to 1. Silhouette scores calculate the quality of clusters and range from −1 to 1, with a value of 1 indicating a high-quality cluster. Metrics were compared across 16S data alone, ANI data, and the novel synteny metric. In addition, due to the sparsity of data in ANI, two other reduced data versions of the synteny metric were used. One contained only values greater than 82% to match the threshold value of ANI, and the other only contained the values for non-zero pairs in the ANI matrix. These were termed "Thr" and "Rem," respectively. The comparison was performed using ANI and synteny coverage in place of the synteny metric, where it was scaled using the covariance metric schema against 16S to calculate the silhouette score.

KNN graphs

K-nearest neighbor (KNN) graphs were constructed for the original 16S distance matrix, the synteny-scaled 16S distance matrix, and the ANI matrix using NetworkX. Using the "distance" mode, the normalized KNN array was visualized as a graph using the Kamada–Kawai Layout of NetworkX. Visualized graphs were labeled with the taxonomic phyla or the taxonomic class. Network quality was calculated using modularity. The networks were compared using (i) Jaccard similarity, (ii) weighted edge Jaccard similarity, (iii) dice coefficients, and (iv) DeltaCon distance (22). The inverse value of the DeltaCon distance

was used to provide similarity. Additionally, DeltaCon node and edge attribution was performed (23). The cluster cutoff was varied to visualize which cutoffs had the highest similarity scores and modularity changes in quality. Community detection of the KNN graphs was performed as a parallel for hierarchical clustering with the Girvan–Newman algorithm, which uses iterative removal of edges based on the shortest path (24). Girvan–Newman communities were displayed as a dendrogram, where the number of communities is chosen by the algorithm.

Functional application

Four cohorts were chosen to apply the synteny measure and augment 16S data. The gene functions of (i) mobile genetic elements, (2) virulence factors, (3) antibiotic resistance, and (4) metabolic genes (consisting of short-chain fatty acids and neurotransmitter genes) were sourced using a combination of pre-existing descriptions in GenBank Flat Files and separate databases. Mobile genetic elements were identified using a set of keywords (transposase, transposon, conjugative, integrase, integron, recombinase, conjugal, mobilization, recombination, and plasmid) from flat file descriptions. Antibiotic resistance nucleotide sequences were gathered from the CARD database and compared to database sequences using BLAST, of which sequences have an identity above 90% (25). The retrieved sequences were BLASTed against the entire database again to identify orthologs, using greater than 95% sequence similarity and a length between 500 and 2,500 base pairs. Virulence factors went through the same process using the Virulence Factor Database (26). The metabolic genes cohort was identified using a combination of gut–brain modules from Vieira-Silva et al. and bile acid metabolism from Funabashi et al., which were sourced as KEGG modules (27, 28). Genes in the KEGG modules that matched the data set's species were downloaded and then BLASTed against the gene database to orthologs once again (29). Each cohort of orthologs was filtered the same way as the original core gene cohort, and synteny similarity was calculated for pairs of genomes that had at least two ortholog pairs. Random forest models were trained on functional synteny to predict taxonomy. The same augmentation process was used to form 16S synteny-scaled distance matrices for each functional cohort, with only values above the 82% threshold. Hierarchical clustering was performed, with silhouette scores and Rand scores calculated. KNN graphs were also visualized with weighted Jaccard and DeltaCon coefficients calculated between the original 16S and original 16S synteny-scaled matrix.

RESULTS

GenBank public data provided genomic, 16S, and core gene data

Our analysis focused on 378 genomes retrieved from GenBank, which represented 10 phyla (*Actinomycetota, Bacteroidota, Campylobacterota, Chlamydiota, Bacillota, Fusobacteriota, Pseudomonadota, Spirochaeota, Mycoplasmatota,* and *Verrucomicrobiota*), 19 classes, 125 genera, and 378 species. *Pseudomonadota* and *Bacillota* had the highest representation (with 135 genomes each), followed by *Actinomycetota* (n = 42) and *Bacteroidota* (n = 29). The other phyla had under 15 genomes per phyla, with *Verrucomicrobiota* having only a single genome. Using the 16S sequences from each genome, we calculated the MASH distance between every pair of genomes, which accounted for multiple unique 16S genes per genome (30, 31). Core genes were identified using the UBCG2 database, resulting in 34,051,278 genes across the database, roughly averaging 70 genes per genome.

Synteny similarity measure indicates genus-level dynamics among bacteria

The developed syntenic measure ranged from 0 to 1, where 1 indicated complete similarity and 0 indicated complete dissimilarity between a given genome pair. The proposed method involves forming multiple rearrangements of the relative gene order, followed by pairwise cosine similarity values for each ortholog pair, and a compilation of data values across the pairs and arrangements (Fig. 1). In the data set used for this study,

synteny values for core genes were found for pairs within the same phyla, representing the phyla of *Bacillota*, *Actinomycetota*, *Pseudomonadota*, *Bacteroidota*, and *Spirochaetota* (Fig. 2A). Smaller subnetworks were identified from the greater network of synteny, where each edge represented a synteny similarity and is weighted by the similarity. The two largest subnetworks were of *Pseudomonadota* and *Bacteroidota*, the largest comprised of the species in *Gammaproteobacteria* class, and the second-largest formed of species in the *Bacteroidales* order (Fig. S6 and S7). *Pseudomonadota*, *Bacteroidota*, and *Bacillota* made up most of the data, with *Actinomycetota* and *Spirochaeta* making up smaller sets. The distribution of synteny similarity values was, therefore, separated into within-genus pairs and out-of-genus pairs. Median synteny similarity values differentiated between pairs in the same genus and pairs outside of the same genus (Fig. 2B). The sparsity of the synteny distance matrix augments 16S distance by using a covariance matrix-based transformation (Fig. 2C). The 16S and synteny distance matrices additionally had low correlation (Spearman = 0.149), indicating different informative values in either matrix. Out of 143,641 total pairwise combinations, there were 553 non-zero values in the sparse similarity matrix and 143,625 non-zero values in the 16S distance matrix. The final scaled matrix has 143,262 non-zero values (0.02% sparsity), increasing the original 16S matrix from 58% sparsity with the synteny matrix of 99% sparsity. The ANI matrix also had 99% sparsity, which indicated that the sparse version of the synteny-scaled metric was necessary for comparison. Two reduced cohorts were formed, labeled "Thr" and "Rem." In Rem, data were removed to only contain the same position values as the ANI matrix (99% sparsity), and in Thr, data were only used at a threshold above 82% similar to ANI field metrics (77% sparsity). Random forest models, which are apt for identifying underlying non-linear trends, were trained to predict

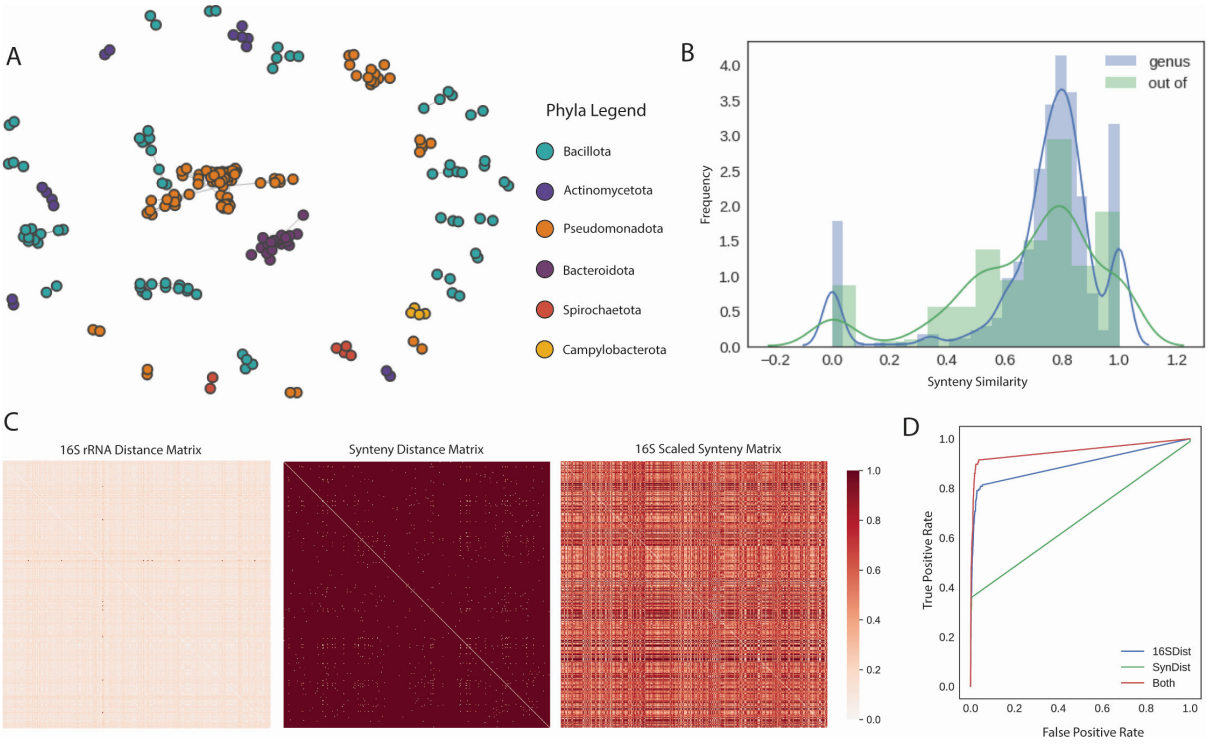

FIG 2 (A) A network diagram showing organization of synteny similarity measure, where nodes represent individual genomes of species and labeled by taxonomic phyla. (B) The count histogram of the synteny similarity measure differentiates between pairs that share the same genus versus pairs that have more taxonomic separation. (C) Each heatmap depicts a particular step of the data analysis, the first being the original 16S distance matrix followed by the synteny distance matrix (calculated as 1—synteny similarity). The last heatmap displayed the final normalized matrix, the combination of the two previous distance matrices. (D) The individual ROC curves for the random forest model trained on taxonomic genus (0, pair in the same genus; 1, pair in the same genus). Each curve represents the data input given (16SDist, 16S MASH distance; SynDist, synteny distance; both, 16S and synteny distance values).

taxonomic distance for data associated with either 16S distance or synteny distance (1—synteny similarity). In all six taxonomic levels, 16S distance performed the best consistently, which lowered at higher taxonomic levels (e.g., phyla and class). For the synteny similarity, only the prediction for the genus level was greater than random, while all others resembled random chance (16S AUC: 0.7229, synteny: 0.649, combined: 0.749) (Fig. 2D).

Hierarchical clustering of 16S synteny-scaled data displays shifting of clusters

Using the 16S synteny-scaled, original 16S, and ANI distance matrices, we performed complete linkage hierarchical clustering. Silhouette scores were used to visualize clustering quality, ranging from +1 to −1 with −1 indicating poor clustering quality and +1 indicating high clustering quality. When varying the cluster cutoff number, the novel synteny metric performed the most consistently for cluster quality via silhouette score in the order of Thr (thresholded metric), Rem (removed metric), and the novel synteny metric termed "Syn" (Fig. 3C). The synteny metric alone without the 16S data performed similarly to the ANI metric as well. The scaling of 16S with ANI data in the same covariance format resulted in negative silhouette scores, indicating reduced clustering assignments (Fig. S5). Synteny coverage also performed negatively compared to 16S and the synteny metric. The resulting dendrograms used a cluster cutoff of 5, where the highest consistent silhouette score was observed, showing differentiated clustering between the scaled and original distance matrices (Fig. 3A and B). The dendrogram from the original matrix clustered *Pseudomonadota* and *Bacteroidota*, while *Actinomycetota* and *Bacillota* were clustered into the same general grouping but

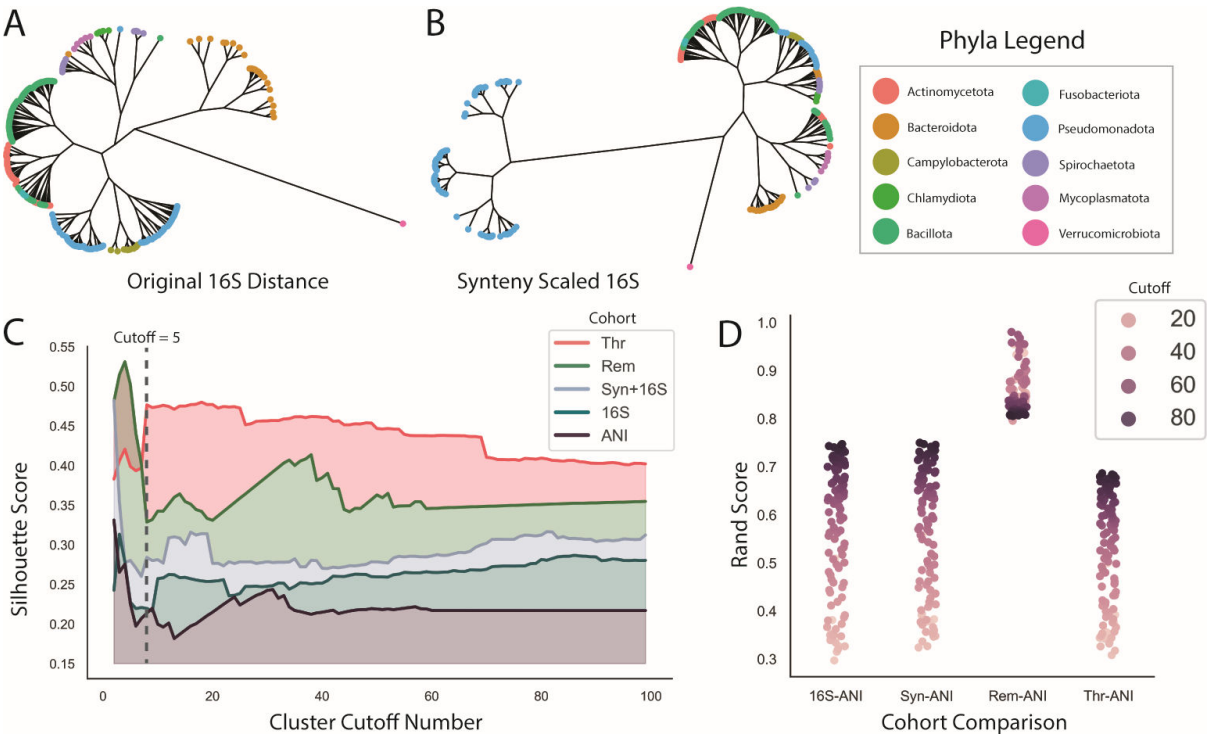

FIG 3 Hierarchical clustering displays increased clustering quality with synteny metric compared to ANI. (A and B) The results of hierarchical clustering on the data are represented as an unrooted dendrogram at a cluster cutoff of 5, with the taxonomic phyla labeled. (A) represents the original 16S data, and (B) represents the synteny-scaled 16S data. (C) Varying the cluster cutoff number from 1 to 100 with the silhouette score across the five different versions of comparisons (Syn+16S, the synteny-scaled 16S data; Thr, the Syn+16S cohort with values only greater than 82% threshold; Rem, the Syn+16S data with the same positional values as the ANI data). (D) Comparisons between each data group and the ANI data, defining the Rand score across the same cluster cutoff modulation. The Rand score shows the similarity of clustering while accounting for possible different cluster numberings. The hue of the point represents the cluster cutoff number.

remained distinct (Fig. 3A). The dendrogram from the scaled matrix slightly shifted the positions of some groupings (Fig. 3B). Most *Pseudomonadota* remained separate, while a small group clustered with *Bacteroidota, Spirochaetota,* and *Chlamydiae* interestingly. The combined clustering of *Bacillota* and *Actinomycetota* remained, with a smaller group cluster further away near *Mycoplasmatota*. *Bacteroidota* also clustered further away. In both dendrograms, *Verrucomicrobiota* clustered the furthest from all other species and remained an outgroup. The remaining data groups of ANI, Thr, and Rem were extremely sparse, resulting in incomplete dendrograms.

In addition, comparisons between clustering results were performed against ANI as a ground truth standard (Fig. 3D). The Rand score measures the similarity of clustering decisions, in which all possible combinations of cluster labels between two sets are considered to account for differential cluster number and labeling. Scores ranged from 0 to 1, with 1 indicating total similarity and 0 indicating no similarity. The highest Rand score is in the Rem-ANI comparison, reflecting the fact that the Rem matrix contains the same position values as the ANI matrix. The 16S and ANI comparison showed the next highest Rand, with the Thr and ANI comparison having a slightly lower Rand score.

KNN graphs offer an alternate form of clustering and visualization of synteny scaling

A secondary approach to analyze differences between the original, synteny-scaled, and ANI data was through KNN graphs. KNN graphs were generated for all distance matrices (16S, SYN, REM, THR, and ANI) across *k* ranging to 15 (Fig. 4D), where a midpoint of $k = 10$ was used for visualization. Nodes indicated species, and edges were weighted by the *k* neighbors algorithm based on the distance values given. Labels were chosen based on taxonomic phyla and class. Visual differences between two networks were based additionally on the layout structure, using the Kamada–Kawai path-length cost function. The phyla-labeled networks show a distinct set of groupings based on phyla (Fig. S8). Similar groups from the hierarchical clustering dendrograms were seen in these networks. In the original 16S network, *Pseudomonadota* separated into two separate groups, whereas in the scaled network, *Pseudomonadota* expanded out linearly (Fig. 4A and B). When the taxonomic labeling was changed to class, this expansion was clearly based on *Gammaproteobacteria*. Many of the other groups remain consistent such as *Bacteroidota, Verrucomicrobiota,* and *Bacillota,* with the *Bacillota* composition splitting into *Bacilli* and *Clostridia* distinctly in both the 16S and scaled KNN graphs. We examined these differences further using the weighted Jaccard and DeltaCon comparative frameworks. The DeltaCon similarity measure (0.928) also assigns attribution to nodes and edges for the impact of the difference between the two networks (Fig. 4C). This edge and node attribution, when accumulated per taxonomic class and normalized, resulted in the highest attribution to *Gammaproteobacteria*, *Bacilli*, *Clostridia*, *Actinomycetia*, and *Spirochaetia*. The lowest values of attribution were seen in *Verrucomicrobiae, Tissierelia, Erysipelotrichia,* and *Deltaproteobacteria*. In contrast, the KNN graph of the whole genome scaled 16S data shows much less consistency of taxonomic groupings other than *Bacillota* and *Pseudomonadota*, wherein the rest of the groups are interspersed. Weighted Jaccard was used for speed to identify similarity between an individual network and the ANI network across the *k*-value modulation (Fig. 4E). Between the four groupings, the 16S-ANI has the highest similarity overall, while the Thr-ANI similarity starts with high similarity at low *k*-values then decreases lower.

Applying syntenic measure to functional gene groups

Four functional gene cohorts were used to test the synteny measure and clustering frameworks, which were antibiotic resistance (ARG), mobile genetic elements (MGE), metabolic genes (MET), and virulence factors (VIR). The number of genes found per genome represents the ubiquity or the specialization of these genes per bacterial group, in which antibiotic resistance and virulence factors tend to be more specific to specific bacterial groups, while metabolic genes and core genes tend to be more universal in

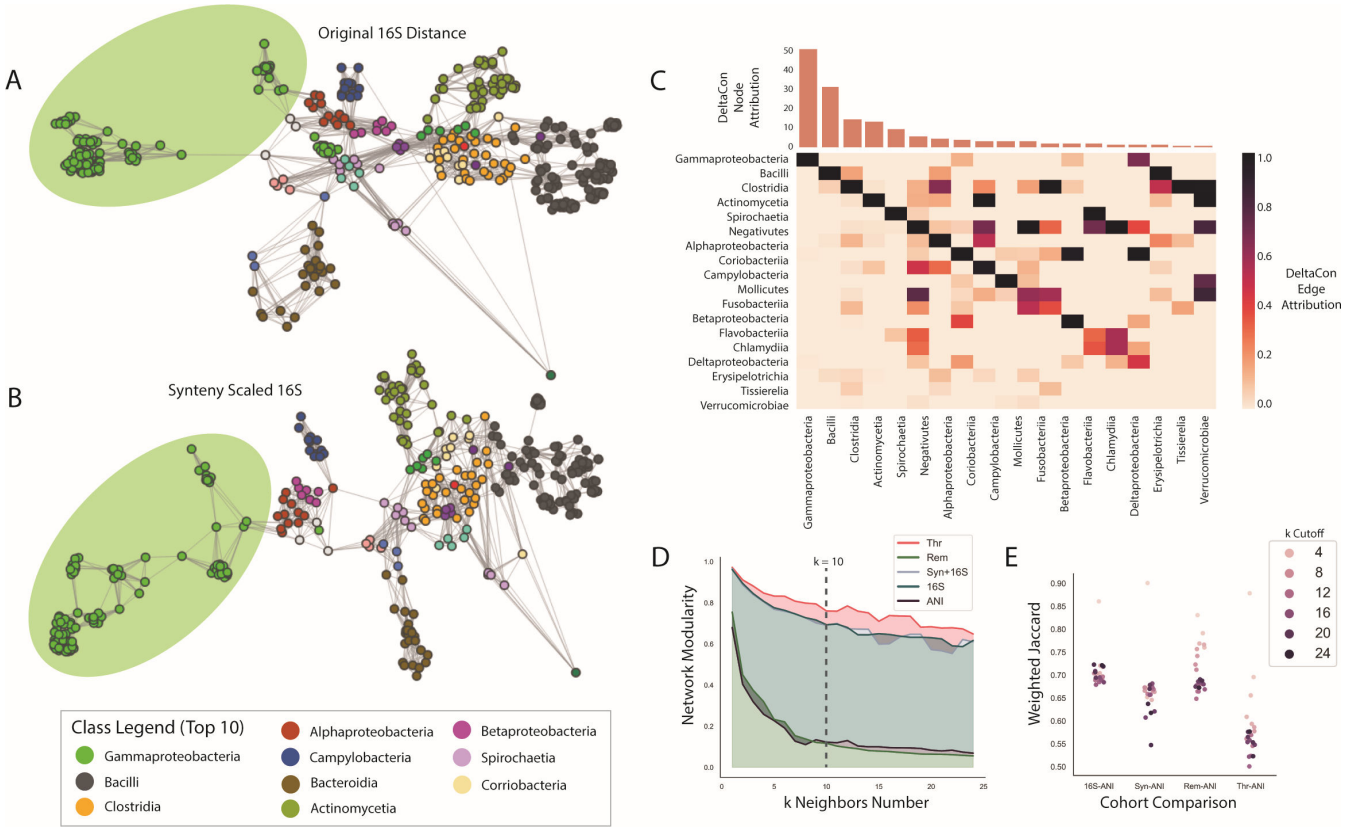

FIG 4 KNN graphs display connection differences of synteny-scaled data. Visualizations of KNN graphs on 16S and synteny-scaled 16S data are displayed in (A and B), labeled based on taxonomic class. (A) represents the original 16S data while (B) displays synteny-scaled 16S data, where only the top 10 classes are displayed in the legend for brevity. The *Gammaproteobacteria* class is highlighted for relevance. (C) contains results of the DeltaCon network comparison method, where nodes and edges are given attribution to quantify their impact on the difference between two networks. Attribution values were accumulated into taxonomic class and then normalized. Node attribution is seen in the top bar graph, ordered based on frequency. Edge attribution is represented by a pairwise heatmap, where higher values indicate that those pairwise edges held more importance to the difference between networks. (D) displays the network modularity over the increasing *k* variable for the KNN graphs. (E) shows the comparisons of paired networks against ANI networks using the weighted Jaccard score.

terms of the numbers of genes present (Fig. 5A). The distribution of synteny similarity measures was quite similar across all five functional groups, of which the original metric based on core genes has the lowest average, while the other four groups have similar averages but different ranges in terms of first and third quartiles (Fig. 5B). The same analyses were performed on the larger cohorts to identify differences between the original synteny metric based on core genes to the functional cohorts in terms of clustering quality and comparison to ANI. Hierarchical clustering modulations showed the highest performance with the ARG group, followed by MET, VIR, MGE, and finally the synteny metric (Fig. 5C). Weighted Jaccard comparisons with the ANI network showed the highest similarity between the MET-ANI, followed by VIR, MGE, and finally ARG (Fig. 5D).

DISCUSSION

The goal of this study was twofold, to (i) suggest an augmentation procedure for 16S rRNA data with a genomic pairwise measure and (ii) provide and test a novel measure based on genomic synteny. Two clustering procedures were used to compare differences in the original and transformed data, one using hierarchical clustering and another using k-nearest neighbor graphs. The ground truth of these clustering approaches ultimately is unknown, particularly given that historical bacterial taxonomic nomenclature has been

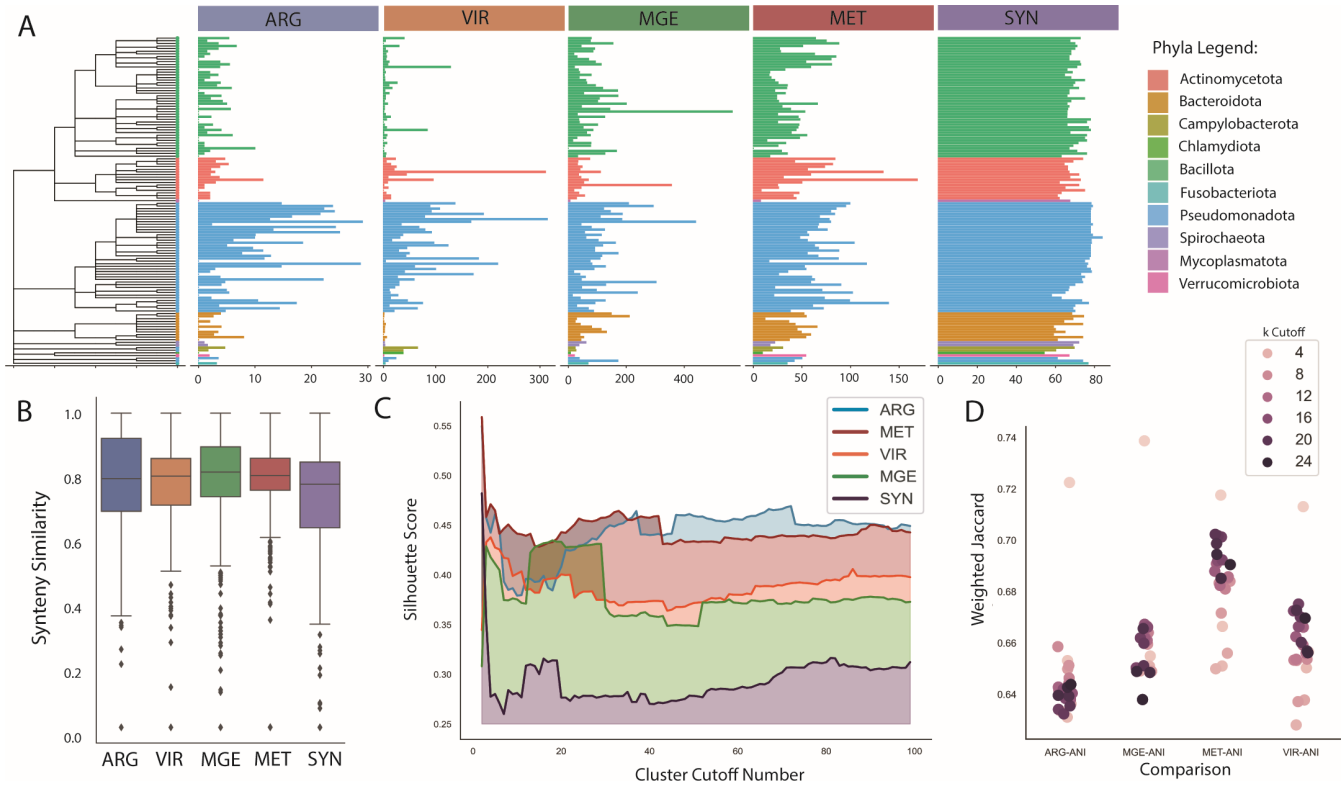

FIG 5 An overall representation of the results of the synteny scaling process on multiple functional gene cohorts. (A) Mean quantities of cohort genes found per genus, where ARG means antibiotic resistance; VIR, virulence factors; MGE, mobile genetic elements; MET, metabolic genes; and COR, core genes. Color labels indicate taxonomic phyla. (B) Synteny similarity distributions as box plots for each functional cohort. (C) Results of hierarchical clustering across cluster cutoff increasing, indicating changes in clustering quality using silhouette scores. (D) Comparisons of cohort KNN graphs against the ANI network using weighted Jaccard, where higher values indicated greater similarity between two networks.

based on limited 16S rRNA values and phenotypic data (32, 33). Therefore, ANI was used as a proxy for ground truth given it is a field standard and consistent against genome fragmentation, to understand how individual versions of the metric fare against ANI.

The results of the synteny similarity measure depict dynamics that match taxonomic relationships. At the genus level, more synteny similarity values are seen within-genus as opposed to out-of-genus (Fig. 2B). Based on the input data for this task (shared core gene orthologs between two genomes), this result validates the assumption that closer related genomes are more likely to share more genes. Therefore, there are more non-zero similarity values present for genomes that are in the same genus as opposed to those outside of the same genus. This also validates the random forest result that prediction only on the genus level is above random. The synteny similarity network also demonstrates that all synteny measures are within-phyla, indicating non-zero values for genome pairs that are more closely related than out-of-phyla pairs (Fig. 2A). The sparsity of the final synteny matrix displays its value when mathematically combined to the original 16S matrix, creating a pairwise matrix that has more variation than the two input matrices (Fig. 2C). The synteny measure on its own, therefore, does not indicate taxonomy other than being within-phyla and genera-specific but is not necessarily predictive of these attributes. Instead, it is most valuable when scaled on other non-sparse data and significantly reduces matrix sparsity.

The hierarchical clustering and KNN graph approaches reveal different pieces of information with respect to the original 16S distance data and the ANI data. In the clustering approach, the quality of the clustering, determined by the silhouette score, is better for the synteny matrices in comparison to the ANI and original 16S matrix. This likely reflects the improvement in the sparsity of the matrix values. In contrast, using the

ANI matrix as an augmentation on 16S resulted in negative silhouette scores but similar Rand scores as 16S and the synteny metric. The Rand scores depict that higher cluster cutoff values lead to better Rand scores or similarity in clusters between ANI and the metric. Interestingly, the opposite occurs in the KNN graphs where lower *k*-values result in higher similarities to the ANI network via the weighted Jaccard metric. Additionally, the Rem matrix performs similarly to ANI likely due to high sparsity of data as well. Therefore, between the sparsity of the Rem matrix and the lower similarities between ANI and the Thr matrix, the original synteny matrix represented the best combination of clustering quality and similarity to ANI for the functional cohorts. These results also highlight an interesting dynamic between the two clustering approaches. Hierarchical clustering draws out general dynamics across the entire data set, particularly stabilizing and improving ANI similarity at higher cluster cutoff values, whereas the KNN approach performs better at low *k*-values, highlighting the strongest pairs. The choice of respective hyperparameters (*k* in KNN and the number of clusters in clustering) determines the interpretation potential and the type of information available in each complementary approach.

The application of these approaches to the functional cohort reveals that the choice of input plays a strong role in the value and interpretation of this metric. While originally core genes were solely used for the metric in the original synteny metric, we observed that antibiotic resistance and metabolism-based genes performed quite well in terms of clustering quality via silhouette scores, with values closer to the Thr matrix in Fig. 3C. In addition, the metabolism cohort also had the highest weighted Jaccard similarity to the ANI networks. While the metabolism cohort has higher amounts of data compared to the core genes, the mobile genetic elements cohort also has higher data quantities yet does not perform as well. Therefore, it is possible that there is some underlying signal where metabolic genes are providing more granular detail on microbial relationships. Previous studies have shown the importance of core metabolic genes as necessary to the minimal bacterial gene set or "pan genome," which is possibly being reflected here (34). However, antibiotic resistance genes also show high clustering quality despite having lower amounts of genes and lower similarity to the ANI networks, whereas the virulence cohort also shows higher similarity to the ANI networks via weighted Jaccard, possibly indicating that the virulence genes represent distances closer to ANI, whereas antibiotic resistance has less similarity. In both approaches, *Gammaproteobacteria* make the highest impact in terms of number of genes per cohort as well as most amount of synteny data, which historically encompass many pathogens as well as have the higher representation of antibiotic resistance and virulence in our data set (Fig. 5A). This is unsurprising given that the data have the most number of genomes for this group of bacteria, potentially due to the clinical bias of GenBank (35). This is particularly visible in the change in the KNN graph structure between the 16S data and synteny-scaled 16S data (Fig. 4A and B). However, despite this bias, there are also an equivalent number of *Bacillota* genomes present in the data and consistent core genes that made up the original synteny metric, which does not significantly change structure. Thus, this phenomenon depicts that as data quality and number of genomes per bacterial class increase, visible changes in clustering and network structure truly represent changes in the genomic data, rather than changing solely as data increases. Metabolic genes are most successful across all groups at replicating ANI relationships and improving clustering quality. In contrast, antibiotic resistance genes can improve novel relationships between bacteria that share resistance, which are less likely to be found by ANI and 16S measures. Thus, it is, therefore, possible to consider this tool in a gene function distribution context as well, to see how functionality affects the final similarity to ANI and whether those functions reflect if a pair of bacteria are known to be related or not.

Current taxonomy and identification improvements have focused on using whole genome alignments instead of 16S sequences, particularly for pathogen outbreak tracking and variant differentiation (36, 37). Some techniques make use of solely

vertically transferred genes, which illustrates a choice of genes for phylogenetics (38). Computational methods have also been designed to combine sequencing data from different PCR-amplified 16S rRNA regions to increase resolution (39). Methods in other species have made use of multiple rRNA sequences (18S, 16S, and 28S) along with cytochrome C to define phylogenetic relationships by combining the alignments of multiple conserved genes (40). The synteny measure also differs from average nucleotide identity by providing detailed information for bacterial relationships that expand a single distance value. Few studies have explored functions as features of genomic relationships but have been explored in the context of horizontal gene transfer (41). As bacterial nomenclature also continues to change and develop with updated information, this augmentation method can provide context-specific relationships that are independent of nomenclature changes (42). We are unaware of other methods that mathematically combine pairwise data with other forms of genomic data or characterize synteny as a pairwise similarity value.

Our proposed method comprises an augmentation approach and a similarity measure that can be applied to pre-existing pairwise data. The choice of genes used for the similarity measure can play a role in determining the strength of relationships between already related bacteria. Some potential use cases for this method can include (i) finding relatives to a novel pathogen based on the synteny of antibiotic resistance or virulence genes, (ii) identifying bacteria with similar horizontal gene transfer profiles based on the syntenic similarity of transferred genes, and (iii) finding the most functionally similar symbionts in different communities that share the same genes and syntenic attributes of those genes. As sequencing technology improves, many outstanding questions remain about how bacterial analysis should be conducted in the future. Is taxonomy relevant for clinical decision-making (43)? Will whole genome sequencing replace 16S for bacterial identification in the future? How can reference-free approaches be considered instead of those that rely on genome references? We suggest that as bacterial genomic data grow in both quantity and quality, augmentation approaches can be applied to understand relationships in context-dependent environments. The proposed synteny similarity measure is one such example that makes use of available data, building on pre-existing knowledge of bacterial taxonomy, and can potentially be applied to functional and clinical contexts.

Conclusion

In this study, we propose a data scaling method, which adds a novel similarity measure to traditional 16S rRNA distance scores, using matrix transformation. Our pairwise measure is based on a graphical structure of a bacterial genome, using ortholog location to form a numerical representation of synteny. Analyzing synteny-scaled 16S rRNA data in comparison to 16S rRNA and ANI data shows that our approach improves on clustering quality, while also retaining ANI relationships, particularly when using metabolic genes as the metric input. In contrast, antibiotic resistance genes can possibly unveil novel relationships that were not previously considered in clinical contexts as well. Ultimately, the choice of clustering method and input gene function determines the interpretation of the relationship between bacteria, making this method a context-aware and dynamic approach that utilizes a novel genomic attribute to determine bacterial relationships.

ACKNOWLEDGMENTS

We thank Dr. Sorin Istrail for his support in developing the graphical structure for synteny similarity. We thank Drs. Tal Korem, Lorin Crawford, Shipra Vaishnava, Peter Belenky, and Rodrigo Bacigalupe for their inputs into the research.

This project was supported in part by Institutional Development Award Number U54GM115677 from the National Institute of General Medical Sciences of the National Institutes of Health, which funds Advance Clinical and Translational Research

(Advance-CTR), as well as the Predoctoral Training Program in Biological Data Science at Brown University from the National Institutes of Health (5T32GM128596-05). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Vivek Ramanan performed the conceptualization, formal analysis, methodology, validation, and writing—original draft. Indra Neil Sarkar performed the conceptualization, methodology, and writing—review and editing.

AUTHOR AFFILIATIONS

[1]Center of Computational Molecular Biology, Brown University, Providence, Rhode Island, USA
[2]Center for Biomedical Informatics, Brown University, Providence, Rhode Island, USA
[3]Rhode Island Quality Institute, Providence, Rhode Island, USA

AUTHOR ORCIDs

Vivek Ramanan http://orcid.org/0000-0002-6406-3938
Indra Neil Sarkar http://orcid.org/0000-0003-2054-7356

FUNDING

Funder	Grant(s)	Author(s)
HHS | NIH | National Institute of General Medical Sciences (NIGMS)	U54GM115677, T32GM128596	Vivek Ramanan Indra Neil Sarkar

DATA AVAILABILITY

The data underlying this article were accessed from the GenBank Public Repository, hosted by the National Center of Biotechnology Information (NCBI). The datasets are available in the article and in its online supplementary material. Sequence data for each cohort will be shared on reasonable request to the corresponding author. The code is publicly available at https://github.com/vivekramanan/synteny-scaling.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Figures (mSystems00497-24-s0001.docx). Figures S1-S9.
Supplemental Tables (mSystems00497-24-s0002.xlsx). Tables S1-S7.

REFERENCES

1. Clarridge JE. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 17:840–862, https://doi.org/10.1128/CMR.17.4.840-862.2004
2. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090. https://doi.org/10.1073/pnas.74.11.5088
3. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data 6:190007. https://doi.org/10.1038/sdata.2019.7
4. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1
5. Hassler HB, Probert B, Moore C, Lawson E, Jackson RW, Russell BT, Richards VP. 2022. Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. Microbiome 10:104. https://doi.org/10.1186/s40168-022-01295-y
6. Rossi-Tamisier M, Benamar S, Raoult D, Fournier P-E. 2015. Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. Int J Syst Evol Microbiol 65:1929–1934. https://doi.org/10.1099/ijs.0.000161
7. Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375. https://doi.org/10.1093/bioinformatics/bty113
8. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x

9. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9
10. Shaw J, Yu YW. 2023. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods 20:1661–1665. https://doi.org/10.1038/s41592-023-02018-3
11. Liu D, Hunt M, Tsai IJ. 2018. Inferring synteny between genome assemblies: a systematic evaluation. BMC Bioinformatics 19:26. https://doi.org/10.1186/s12859-018-2026-4
12. Fitzgerald J, Freeman T, Bach B, Harling-Lee JD. 2023. Visualising the bacterial pangenome: an analysis of the genetic content of *Staphylococcus aureus*. https://doi.org/10.7488/era/3166
13. Roja B, Saranya S, Chellapandi P. 2023. Discovery of novel virulence mechanisms in *Clostridium botulinum* type A3 using genome-wide analysis. Gene 869:147402. https://doi.org/10.1016/j.gene.2023.147402
14. Xiao Z, Lam HM. 2022. ShinySyn: a Shiny/R application for the interactive visualization and integration of macro- and micro-synteny data. Bioinformatics 38:4406–4408. https://doi.org/10.1093/bioinformatics/btac503
15. Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. 2013. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In Algorithms in bioinformatics. WABI 2013, Lecture notes in computer science.
16. Lee J, Hong W-Y, Cho M, Sim M, Lee D, Ko Y, Kim J. 2016. Synteny portal: a web-based application portal for synteny block analysis. Nucleic Acids Res 44:W35–W40. https://doi.org/10.1093/nar/gkw310
17. Blin K. 2021. NCBI-genome-download. Available from: https://github.com/kblin/ncbi-genome-download#readme
18. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114
19. Kim J, Na S-I, Kim D, Chun J. 2021. UBCG2: up-to-date bacterial core genes and pipeline for phylogenomic analysis. J Microbiol 59:609–615. https://doi.org/10.1007/s12275-021-1231-4
20. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219
21. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69:e96. https://doi.org/10.1002/cpbi.96
22. Tantardini M, Ieva F, Tajoli L, Piccardi C. 2019. Comparing methods for comparing networks. Sci Rep 9:17557. https://doi.org/10.1038/s41598-019-53708-y
23. Koutra D, Shah N, Vogelstein JT, Gallagher B, Faloutsos C. 2016. DeltaCon: principled massive-graph similarity function with attribution. ACM Trans Knowl Discov Data 10:1–43. https://doi.org/10.1145/2824443
24. Despalatovic L, Vojkovic T, Vukicevic D. "Community structure in networks: Girvan-Newman algorithm improvement." 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); Opatija, Croatia: p 997–1002. https://doi.org/10.1109/MIPRO.2014.6859714
25. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, et al. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 48:D517–D525. https://doi.org/10.1093/nar/gkz935
26. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res 44:D694–D697. https://doi.org/10.1093/nar/gkv1239
27. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, Vandeputte D, Valles-Colomer M, Hildebrand F, Chaffron S, Raes J. 2016. Species-function relationships shape ecological properties of the human gut microbiome. Nat Microbiol 1:16088. https://doi.org/10.1038/nmicrobiol.2016.88
28. Funabashi M, Grove TL, Wang M, Varma Y, McFadden ME, Brown LC, Guo C, Higginbottom S, Almo SC, Fischbach MA. 2020. A metabolic pathway for bile acid dehydroxylation by the gut microbiome. Nature 582:566–570. https://doi.org/10.1038/s41586-020-2396-4
29. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. 2023. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 51:D587–D592. https://doi.org/10.1093/nar/gkac963
30. Větrovský T, Baldrian P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 8:e57923. https://doi.org/10.1371/journal.pone.0057923
31. Drew J, Hahsler M. 2014. "Strand: fast sequence comparison using mapreduce and locality sensitive hashing" Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics
32. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–2764. https://doi.org/10.1128/JCM.01228-07
33. Yutin N, Galperin MY. 2013. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. Environ Microbiol 15:2631–2641. https://doi.org/10.1111/1462-2920.12173
34. Gil R, Silva FJ, Peretó J, Moya A. 2004. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68:518–537. https://doi.org/10.1128/MMBR.68.3.518-537.2004
35. Ramanan V, Mechery S, Sarkar IN. 2022. Genbank as a source to monitor and analyze host-microbiome data. Bioinformatics 38:4172–4177. https://doi.org/10.1093/bioinformatics/btac487
36. Maguvu TE, Bezuidenhout CC. 2021. Whole genome sequencing based taxonomic classification, and comparative genomic analysis of potentially human pathogenic *Enterobacter* spp. isolated from chlorinated wastewater in the North West province, South Africa. Microorganisms 9:1928. https://doi.org/10.3390/microorganisms9091928
37. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. Clin Microbiol Rev 30:1015–1063. https://doi.org/10.1128/CMR.00016-17
38. Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM. 2021. Prokaryotic taxonomy and nomenclature in the age of big sequence data. ISME J 15:1879–1892. https://doi.org/10.1038/s41396-021-00941-x
39. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. 2018. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 6:17. https://doi.org/10.1186/s40168-017-0396-x
40. Edgecombe GD, Giribet G. 2004. Adding mitochondrial sequence data (16S rRNA and cytochrome *c* oxidase subunit I) to the phylogeny of centipedes (Myriapoda: Chilopoda): an analysis of morphology and four molecular loci. J Zool Syst Evol Res 42:89–134. https://doi.org/10.1111/j.1439-0469.2004.00245.x
41. Zhou H, Beltrán JF, Brito IL. 2021. Functions predict horizontal gene transfer and the emergence of antibiotic resistance. Sci Adv 7:eabj5056. https://doi.org/10.1126/sciadv.abj5056
42. Carroll KC, Munson E, Butler-Wu SM, Patrick S. 2023. Point-counterpoint: what's in a name? Clinical microbiology laboratories should use nomenclature based on current taxonomy. J Clin Microbiol 61:e0173222. https://doi.org/10.1128/jcm.01732-22
43. Janda JM. 2018. Clinical decisions: how relevant is modern bacterial taxonomy for clinical microbiologists? Clin Microbiol Newsl 40:51–57. https://doi.org/10.1016/j.clinmicnews.2018.03.005