Bioinformatics, 38(17), 2022, 4172–4177
https://doi.org/10.1093/bioinformatics/btac487
Advance Access Publication Date: 8 July 2022
Original Paper

Data and text mining

GenBank as a source to monitor and analyze Host-Microbiome data

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

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

*To whom correspondence should be addressed.

Associate Editor: Zhiyong Lu

Received on March 10, 2022; revised on June 8, 2022; editorial decision on June 28, 2022; accepted on July 7, 2022

Abstract

Motivation: Microbiome datasets are often constrained by sequencing limitations. GenBank is the largest collection of publicly available DNA sequences, which is maintained by the National Center of Biotechnology Information (NCBI). The metadata of GenBank records are a largely understudied resource and may be uniquely leveraged to access the sum of prior studies focused on microbiome composition. Here, we developed a computational pipeline to analyze GenBank metadata, containing data on hosts, microorganisms and their place of origin. This work provides the first opportunity to leverage the totality of GenBank to shed light on compositional data practices that shape how microbiome datasets are formed as well as examine host–microbiome relationships.

Results: The collected dataset contains multiple kingdoms of microorganisms, consisting of bacteria, viruses, archaea, protozoa, fungi, and invertebrate parasites, and hosts of multiple taxonomical classes, including mammals, birds and fish. A human data subset of this dataset provides insights to gaps in current microbiome data collection, which is biased towards clinically relevant pathogens. Clustering and phylogenic analysis reveals the potential to use these data to model host taxonomy and evolution, revealing groupings formed by host diet, environment and coevolution.

Availability and implementation: GenBank Host-Microbiome Pipeline is available at https://github.com/bcbi/genbank_holobiome. The GenBank loader is available at https://github.com/bcbi/genbank_loader.

Contact: neil_sarkar@brown.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Started in 1982, the GenBank molecular sequence database has had a significant role in the growth and increase of genomic research. GenBank is maintained by the National Center of Biotechnology Information (NCBI) at the US National Library of Medicine and is often a main reference for genomic analysis. As of this writing, GenBank contains 476 million records, with 223 542 prokaryotic genomes representing 14 606 species, and 1290 eukaryotic species genomes (Sayers *et al.*, 2022). GenBank is populated through a combination of submissions from individual researchers, collaborative consortia, as well as internal NCBI efforts. GenBank supports molecular sequence studies across many fields, including population genetics, disease analysis and the microbiome (Cho *et al.*, 2000; Powell *et al.*, 1995; Sarkar, 2010). GenBank data compositional studies to date have been done on molecular sequence data and not specifically in a microbiome context.

The microbiome, the combination of all the microorganisms in a host, has been associated with numerous disease and functional phenotypes in humans (Cho and Blaser, 2012). Studies have examined the microbiome's role in the gut–brain axis, cancer risk, digestion, obesity, immune modulation, drug metabolism, as well as risk for disease or dysfunction (Blacher *et al.*, 2017; Cryan and O'Mahony, 2011; Helmink *et al.*, 2019; Lawrence and Hyde, 2017; Ley, 2010; Zimmermann *et al.*, 2019). Microbiome analyses rely on the identification of the microorganisms within a given sample. Species identification of microorganisms with a sample is typically done using sequencing, which includes the alignment of molecular sequence data to reference or representative genomes (Malla *et al.*, 2018). This compositional analysis relies on knowing which references are available from common molecular sequence databases, such as GenBank. Understanding the taxonomic composition of GenBank is therefore an essential step to advance microbiome research.

This study is the first to retrospectively analyze GenBank data composition for the microbiome. Previous studies have examined the utility of GenBank metadata for a range of other contexts, including those that examine comparative biology or phylogeography (Chen and Sarkar, 2010, 2011; Magge *et al.*, 2020; Scotch *et al.*, 2011; Tahsin *et al.*, 2016; Weissenbacher *et al.*, 2015). Through a novel panel of methods to analyze GenBank data, we

assessed the potential of GenBank to support microbiome studies. The data composition of GenBank can be best assessed by analyzing its metadata. Each GenBank microorganism submission or record, aside from its genomic information and sequence, contains the information of where and how it was collected and identified. In the context of host–microbiome organisms, the metadata indicate the species of the host and the tissues or organs that they were cultured or found in. The combination of data on the microorganism type, the host, as well as the tissue origin of the microorganism, provide ample data to analyze host–microbiome relationships across specific sites. For this study, we focused on analysis of gut microbiome data, which form the dominant source sites in GenBank. Specifically, we analyzed the data composition of microbiome data across multiple hosts from the animal kingdom. We used an unsupervised clustering technique to identify clusters of hosts based on microbiome similarity as well as generated a phylogenetic tree of the hosts using microbiome composition to model host–microbiome coevolution.

Harnessing repositories such as GenBank for large-scale retrospective studies, such as the study presented here, requires the development of methods to standardize data originating from disparate studies and naming schemas. Contributions and insights of this study include (i) displaying the pathogenic focus within GenBank microbiome data; (ii) exploring host–microbiome data in the context of evolution and clustering; and (iii) demonstrating the potential for retrospective GenBank data outside of representative sequences and as a dataset on its own. This study makes use of GenBank as a nontraditional microbiome dataset and demonstrates the potential value in its use to understand pathogenic microbiomes, particularly in the context of host–microbiome evolution.

2 Material and methods

2.1 Collecting and processing data from GenBank

Microbiome and associated host information was imputed from GenBank metadata, which were systematically downloaded and extracted using an updated version of a Java program (genbank-loader: https://bitbucket.org/UVM-BIRD/genbank-loader/src/master/). The tool retrieved and extracted GenBank metadata into a MySQL database, which included a record of a given microorganism, along with its provided source site, the recorded host and the tissue type. The data from the fields were mapped to standardized terms. Data processing was done in Python3 and Julia 1.5.4. Isolation source and tissue type data were then mapped to Unified Medical Language System (UMLS) concepts using the MetaMap tool from the National Library of Medicine (Aronson, 2001), merged into a single dataset including only isolation sources that were tissue types or organs, and then categorized into general groupings for tissue type source sites. Because of variation among the manually input data and terminology across different types of host animals, standardized groupings were created to represent terms across a known microbiome region. The groupings of interest for this study were gastrointestinal (GI) concepts (keywords: feces, abdomen, intestine, rectum, stomach, esophagus, rumen, colon, ileum, cecum, jejunum and duodenum).

Host data were processed using a combination of MetaMap and the NCBI Entrez Taxonomy database to map GenBank record metadata to scientific names of animal hosts. For this study, only hosts with species-level scientific names were kept (i.e. discarding hosts with vernacular or genus-level names). Additionally, only hosts with GI tracts were kept from the following classes: Mammalia, Aves, Reptilia, Amphibia, Cephalopoda and Actinopterygii. Microbe organism data were mapped to NCBI Entrez Taxonomy database and used to retain bacteria, archaea, fungi, protozoa, viruses and invertebrate parasites. Uncultured or unspecified microorganisms of nonspecific phyla were not included.

The resultant combined data matrix was organized into a presence–absence matrix, where each host had a binary value indicating whether a particular phylum of microorganisms was found in the host data. Presence was indicated by a '1' and absence indicated by a '0', with all microorganismal phyla being given equal weight. Non-microbiota phyla were filtered to retain only phyla with at least one data point amongst all the included hosts.

For the host data, weighted sampling was used to equalize coverage amongst hosts that were overrepresented. Only hosts with at least four presence values were kept for the matrix. If the host data consisted of more than 10 phyla, then a weighted sampling was performed. The probability distribution of the weighted sampling was calculated by measuring the percentage of records in the dataset for a microorganism per host. Each phylum was redrawn based on the weighted probability distribution to create a coverage corrected host data.

2.2 Composition of GenBank data

Composition analysis of GenBank data was performed in Python with Matplotlib, using Jupyter Notebooks. Percentages were calculated by tabulating the number of unique species present in every phylum for a given grouping across all hosts as well as for humans. An incidence-specific Hill Numbers approach, extended by Chao *et al.*, was used to quantify diversity values and create rarefaction curves, using a combination of $q = 0$, $q = 1$ and $q = 2$ Hill Numbers (Chao *et al.*, 2014). Sample cohorts were randomly sampled from the original presence–absence matrix from the GI data at both the phylum and species level without replacement.

2.3 Unsupervised clustering using markov clustering

The Markov Clustering Algorithm (MCL) was used to form clusters of hosts using unsupervised learning. MCL requires a pre-formed network, with edges connecting nodes based on similarity. The network for this study consisted of nodes representing hosts, with edges based on cosine similarity. Cosine similarity measures the cosine of the angle of difference between two vectors, where each vector represents a host's presence–absence data.

For vectors A and B with presence–absence data, cosine similarity ranges from 0 to 1, with 1 being the most similar and 0 being the least similar. Cosine similarity values were calculated for all pairwise combinations of hosts based on their microbiome compositional data. Edges were added between nodes for cosine similarity values >0.66, which approximately indicated an angle of difference less than 50 degrees. Networks were created for all hosts, as well as solely mammalian hosts, the latter of which was only used for visualization. The hosts were then clustered using the MCL algorithm. For visualization, each node was labeled with the cluster and taxonomical data. The clustering results were visualized in Cytoscape, where node size was dependent on the degree of associated edges. Cluster composition was analyzed by calculating the percentage present per each phylum shared across cluster members. Cluster composition was visualized as a heatmap using Seaborn in Python.

2.4 Analyzing host evolution using microbiome composition

A Neighbor Joining phenetic (distance-based) tree was created from the presence–absence host data using PAUP* (Phylogenetic Analysis Using Parsimony*and other methods). The tree was rooted at *Octopus mimus*, which was identified as the most ancestral host in the dataset based on TimeTree (http://timetree.org; Kumar *et al.*, 2017). The tree and taxonomic class groupings were visualized using ggtree and treeio in R (Wang *et al.*, 2020; Yu, 2020). A reference evolutionary tree was created in TimeTree using the same hosts from the microbiome tree, rooted at *Octopus vulgaris*, and visualized in R with ggtree. Hosts that were not found in the TimeTree database were substituted based on TimeTree host naming conventions.

3 Results

3.1 Outputs of the GenBank database pipeline

In total, the GenBank loader program retrieved and standardized 184 659 278 records. Of the 163 905 unique microbial species total in the database, 9070 could not be identified and 885 uncultured

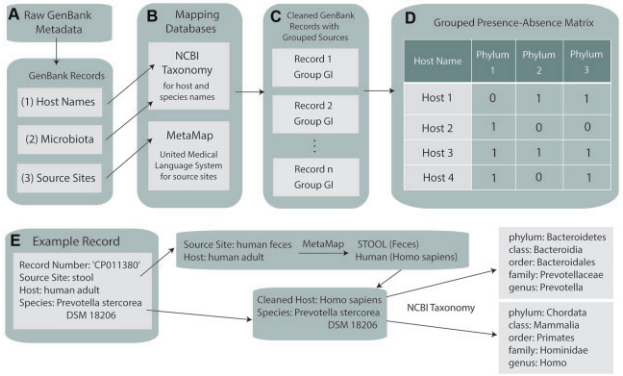

Fig. 1. Visualization of the GenBank Host-Microbiome Pipeline steps. After initial download of GenBank Metadata, data are organized into three data types (**A**) after which it is mapped to two different databases (**B**) to clean up manual input errors and classify groupings. The cleaned records (**C**) are grouped together based on a chosen group and turned into a presence–absence matrix (**D**) based on each host–microorganism record, indicating a presence. (**E**) An example record going through Step B, in which the source site and host name are standardized using MetaMap, after which the cleaned host name and species name are referenced to NCBI Taxonomy for taxonomic information.

specimens were removed. The GI grouping consisted of 37 007 067 records, the largest of any of the source sites. The corresponding GI presence–absence matrix had 132 hosts. Genus level hosts were excluded and host names with the same scientific name, but different naming conventions were merged. Additionally, the 59 microbial phyla, consisting of 19 bacterial, 10 protistan, 17 viral, 3 invertebrate, 9 fungal and 1 bacteriophage phyla, were excluded as hosts.

3.2 Insights from the GenBank data

The 132-host presence–absence matrix contained 28 373 microbial species. The general composition of the microbial phyla consisted of *Proteobacteria* at 30%, *Firmicutes* at 23%, *Pisuviricota* at 9%, *Actinobacteria* at 7%, *Bacteroidetes* at 5%, *Apicomplexa* at 4%, *Ascomycota* at 3% and smaller phyla under 2% (Fig. 1A). This composition was observed in other groupings as well, specifically mammals and humans (Fig. 1B and C). Based on this data skew towards species in the *Proteobacteria* phylum, the rest of the analysis was performed on the phylum level to reduce bias towards *Proteobacteria* species data. Additional analysis of other mammal groupings, specifically primates and bats, indicated other forms of data bias. Primates were dominated by *Apicomplexa* followed by viral phyla, while bats only had viral phyla present (Supplementary Fig. 1).

The biodiversity estimation based on Hill Numbers revealed the impact of human data on the biodiversity of the dataset (Fig. 2C). When human data were introduced into biodiversity calculations, all three metrics (species richness, Shannon diversity and Simpson diversity) showed sharp increases in diversity values. Analysis at the phylum level showed a balanced approach to biodiversity estimation, where no host drastically increased the estimator's values. The phylum level was used for all analysis because it did not bias towards any hosts that drastically impacted biodiversity estimates. On the phylum level, *Proteobacteria* had a presence value in 64% of the hosts, followed by *Pisuviricota*, *Apicomplexa* and *Firmicutes* being present in approximately half of the hosts. The presence data from the GenBank GI dataset was visualized as a network, where each edge modeled a presence value between host and microorganism phyla (Fig. 3).

3.3 Clustering of GenBank GI data forms unsupervised groups based on microbiome composition

The Markov Clustering algorithm (MCL) resulted in the formation of 16 clusters using phylum level GI data. The four clusters that contained most of the hosts were Clusters 1, 2, 3 and 6 (Fig. 4A). Clusters 1 and 2 had mixtures of Mammalia (58% of hosts in Cluster 1 and 76% of hosts in Cluster 2) and Aves (42% hosts in

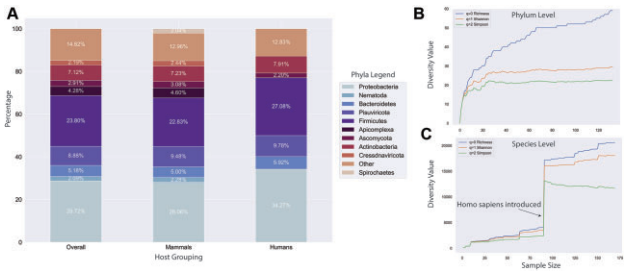

Fig. 2. Species based data composition and diversity analysis. (**A**) Percentage compositional makeup per phyla of entire dataset (Overall), mammalian subset (Mammals) and human data (Humans). Each percentage indicates the number of unique species in each phylum, averaged across the total number of species. All phyla under 2% representation were merged into a general group, 'Other'. *Proteobacteria* and *Firmicutes* dominate all three datasets based on unique species found. Panels (**B**) and (**C**) depict biodiversity estimator analysis using q-based Hill Numbers. (B) Hill numbers approach for incidence data at the phylum level. (C) Hill numbers approach for incidence data at the species level. Sample size refers to number of hosts used per run and diversity value reflects the estimator output for each *q*-value. The large increase in diversity value in (C) occurs at the introduction of Homo sapiens data into the analysis.

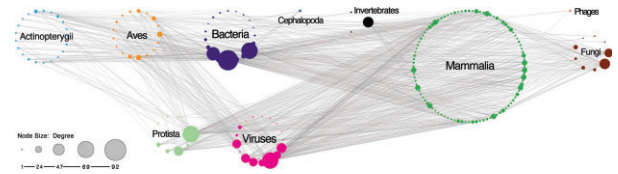

Fig. 3. Visualization of the GI presence–absence data matrix, grouped based on taxonomical class for hosts and domain for microorganism phyla. Visualization was processed in Cytoscape, in which node size is representative of the degree of edges for the node. Edges indicate a presence value between a host–microorganism phyla pair. Each host class and microorganism phyla domain are colored individually and labeled distinctly.

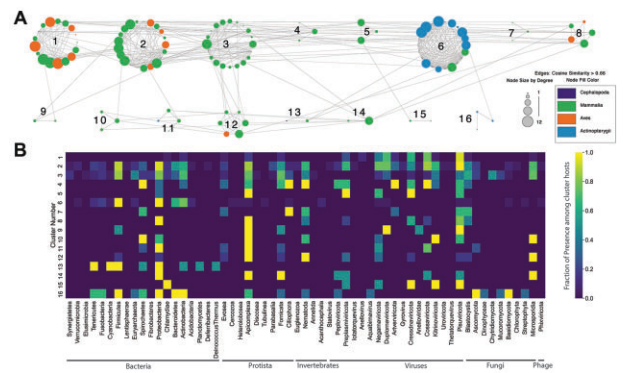

Fig. 4. Visualization of results from Markov Clustering Linkage (MCL) algorithm. (**A**) Grouped clusters from GI phylum data, color corresponding to taxonomical class and size of nodes reflecting degree per node. Clusters labeled based on unsupervised ordering of the algorithm. (**B**) Microorganism phyla composition per cluster, calculated as fraction of presence per each phylum among each cluster's hosts. Phyla are grouped based on microorganism domain accordingly.

Cluster 1 and 24% hosts in Cluster 2). Cluster 3 contained only mammals, of which 50% were Artiodactyla or even-toed ungulates. Cluster 6 contained most of the Actinopterygii (76% of hosts in Cluster 6), followed by 24% mammalian hosts. The smaller clusters contained groupings of mammals and Aves or mammals alone. Clusters 4 and 7 solely had primates, while Cluster 9 consistent of only rodents. Certain phyla were predominantly shared by hosts of specific clusters, which were visualized as a heatmap (Fig. 4B). The largest four clusters (Clusters 1, 2, 3 and 6) had high percentages of

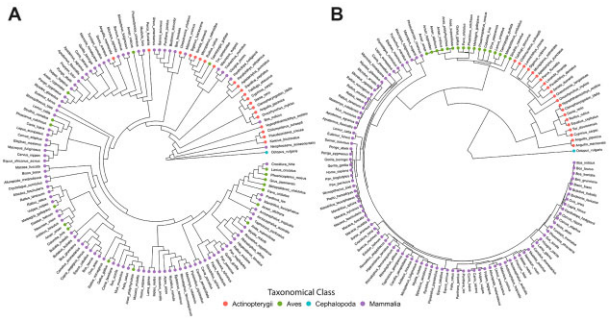

Fig. 5. Evolutionary trees of microbiome composition and true evolutionary tree. (**A**) Microbiome based evolutionary tree, created with PAUP* and labeled based on taxonomical class of the host. (**B**) True evolutionary tree based on phylogenetic distance, created with TimeTree, and using same hosts from Tree A with substitutions based on database labeling differences. Both trees visualized with ggTree in R and rooted with oldest evolutionary host, *Octopus vulgaris*.

Proteobacteria and *Firmicutes*. Cluster 1 was most defined by viral phyla such as *Pisuviricota* and *Cossaviricota*, while Cluster 2 showed large range of shared phyla across all kingdoms. Cluster 3 was defined by the dominant bacterial phyla, as well as *Apicomplexa* and *Pisuviricota*, while Cluster 6 contained *Bacteroidetes* alongside the dominant bacterial phyla.

3.4 Evolutionary analysis of GenBank GI data models host–microbiome coevolution

The evolutinary trees created from the GI presence–absence data identified some evolutionary supported groupings and other unsupported clades (Fig. 5A). Humans were distinct from most primates, alongside domesticated animals, animal model organisms, rhesus macaques and livestock animals. Other mammals fell into two groupings. One mammal group consisted of larger mammals (e.g. buffalos and camels). The second mammal group consisted of a mixed smaller group (e.g. lemurs and deer), indicating more shared attributes. Ray-finned fishes (Actinopterygii) were grouped in proximity and closer to the root of the tree. Birds (Aves) were distributed across the clades. Rodents were also observed to be grouped together. Primates were mostly found together, noticeably missing humans and rhesus macaques.

Compared to a reference evolutionary tree (Fig. 5B), the positioning of the Actinopterygii was the main consistent grouping across both trees. The differences in evolutionary distances between clades and common ancestor branching of the clades were also visible between the reference evolutionary tree and the microbiome based generated tree.

4 Discussion

Large and publicly accessible repositories like GenBank can be rich sources for large-scale retrospective secondary analyses, which would be cost and labor prohibitive as primary prospective studies. However, for the data transformation steps needed for this study, there is a necessity to choose methods and data representations that accommodate the realities of collection focused repositories like GenBank. The core of the analysis for this study focused on the development and use of a presence–absence matrix. This was because in GenBank the absence of a microorganism for a given host did not necessarily indicate that the microorganism did not exist in that host. Instead, it simply indicates that there is no record of that microorganism found or identified within the given host. In contrast to the typical use of abundance data in microbiome studies, the number of times a given microorganism is found associated with a host in the GenBank data is not indicative of its abundance. A presence–absence matrix only accounts for host–microbiome pairs and does not make any implications on abundance. A presence–absence matrix further removes the necessity of data normalization within a particular phylum. Regardless of the dominance of a species within a given phylum, a single species within the phyla will result in a presence value for the phyla. In addition, when considering diversity estimation of the GenBank dataset, alpha-diversity, the traditional microbiome diversity measure, cannot be performed. This is because of the lack of abundance data to calculate evenness and richness, as well as lack of phylogenetic distances between bacterial and viral phyla. Instead, an incidence-specific Hill Numbers approach was used to approximate biodiversity richness. This reasoning extends to the calculation of beta-diversity, the traditional pairwise difference method for microbiome samples. Cosine similarity was chosen as the similarity metric because of its effectiveness on presence–absence data. Other, more traditional, microbiome beta-diversity measures require microbiome abundance data or are limited to bacterial data. Unweighted UniFrac was considered because of the use of presence–absence data and the addition of phylogenetic components in the diversity metric. However, evolutionary distances of viral phyla in relation to bacterial, archaeal, and eukaryotic phyla could not be determined. The cosine similarity threshold of 0.66 was chosen based on the distribution of cosine similarity values and network size. A threshold of 0.5 resulted in 17% of all pairwise connections included in the network, while 0.75 only included 1% of all pairwise connections. Therefore, 0.66 or 2/3 was used, which included 5% of all pairwise connections.

The Markov Clustering Linkage algorithm (MCL) was chosen for this study to form clusters out of a pre-formed network based on cosine similarity. There are myriad unsupervised clustering techniques such as K-means or graph-based clustering algorithms (Brohee and van Helden, 2006; Shi *et al.*, 2022). MCL was chosen because it is unsupervised and non-parametric, therefore inferring the number of clusters with high computational efficiency compared to other graph-based clustering algorithms (Azad *et al.*, 2018). Finally, the construction of the microbiome phylogenetic tree was performed using distance metrics, rather than character-based tree algorithms. The neighbor joining algorithm was chosen because the distance between any two taxa is calculated based solely on the presence values. Other tree inferencing algorithms, such as maximum parsimony or maximum likelihood, depend on a chosen model of evolution or the assumption that absence data indicates true absence, which were not possible for this study.

The results of this study emphasize the importance of compositional analysis for host–microbiome data in GenBank. Analysis of the GI presence–absence matrix reveals a notable bias of the dataset towards *Proteobacteria*, which consists of many pathogenic species (Rizzatti *et al.*, 2017; Shin *et al.*, 2015). Indeed, the genomic information of microbial pathogens are used to understand disease, improve clinical practice, and create new therapies, therefore increasing the likelihood of pathogenic organisms included in GenBank. However, human gut microbiome composition is known to be comprised of 90% *Bacteroidetes* and *Firmicutes*, the ratio of which has been associated with gut dysbiosis and aging (Magne *et al.*, 2020; Mariat *et al.*, 2009). In contrast, the GenBank human GI data shows higher composition of *Proteobacteria* and *Firmicutes*, with *Bacteroidetes* making up a much smaller percentage (Fig. 2C). *Bacteroidetes* are well-associated with humans, with the genera *Prevotella* associated with fiber heavy diets and *Bacteroides* as an opportunistic pathogen in westernized microbiomes (Chen *et al.*, 2017; Vangay *et al.*, 2018). It is possible that the decreased *Bacteroidetes* data is due to challenges in culturing anaerobic species. Additionally, diseased patients often have increased levels of *Proteobacteria* and decreased *Bacteroidetes* during analysis of their gut microbiomes, thus it is possible that microbial composition in GenBank is reflective of a diseased human gut (D'Argenio and Salvatore, 2015). However, given that GenBank data represents the record collection of multiple studies on the microbiome as well as the specification of phylum-level analysis in this study, the ratio of *Proteobacteria* to *Bacteroidetes* could be due to a combination of both a clinical pathogenic focus as well as sampling diseased patients through that context. Another area of low representation is Archaea. The phylum *Euryarchaeota* plays an important role in the gut microbiome and is normally characterized as one component of microbiome composition. However, there is likely limited data on

this phylum because Archaea are hard to culture and identify (Eckburg *et al.*, 2005). The remaining phyla are comprised of fungi, protozoa, viruses, and invertebrates, of which fungi and protista have been linked to digestive functions while viruses and invertebrate worms can impact microbial diversity (Chabe *et al.*, 2017; Firkins *et al.*, 2020; Garmaeva *et al.*, 2019; Gilbert *et al.*, 2021; Gouba and Drancourt, 2015; Mukhopadhya *et al.*, 2019).

The hosts identified in this study skew towards mammals, of which humans, livestock, and model organisms comprise the majority. While host organisms from the Aves and Actinopterygii classes were also included in this study, some taxonomical classes (e.g. Reptilia and Amphibia) were not included because of lack of data (i.e. fewer than four microbial species associated species) or challenges with standardization (e.g. ambiguous genus or common names). Mammalian analysis alone indicated higher data coverage in the following orders: Carnivora, Rodentia, Primates, Artiodactyla and Chiroptera (Supplementary Fig. 2). Most of the hosts with higher coverage included model organisms used in scientific research and livestock animals, often studied for the food industry. The high amounts of livestock animals indicate that the effects of domestication may impact microbiome analysis as well, yielding differences between wild and captive/domesticated animals (Prabhu *et al.*, 2020). Domestication impacts on animals have been paralleled with the impact of industrialization on humans (Reese *et al.*, 2021). Plant microbiota also shows reduced diversity under domestication, affecting crop production (Martinez-Romero *et al.*, 2020). Because domestication is a nascent area of study with the gut microbiome, there are limited studies examining its impact on microbiome–host coevolution. There may be opportunities to further examine the impact of domestication on microbiome composition given the high amount of domesticated host animals in GenBank.

The results of Markov Clustering Algorithm (MCL) demonstrated how hosts could be grouped based on similarities in microbiome composition (Fig. 3A). The resulting groupings reflected the ubiquity of certain microorganism phyla in determining cluster composition (Fig. 3B). It also delineated rough groupings of host taxonomy, in which Actinopterygii separated from other Mammalia and Aves groupings. This finding was also observed in the results of the microbiome-based evolutionary tree (Fig. 5A). This could be due to a multitude of reasons. Host diet could play a role in affecting microbiome composition of different host animals. Gut microbiome functions of herbivorous and carnivorous mammals show differences in microbiome composition based on diet. For example, herbivores tend to have more carbohydrate-metabolizing symbionts and carnivores have more protein-metabolizing enzymes (Muegge *et al.*, 2011). Dietary impacts on gut microbiomes have been tested with studies on human diets impacting microbiome composition (Singh *et al.*, 2017). High animal protein diets show increased bile-tolerant bacteria (*Proteobacteria*, *Bacteroidetes*) and plant protein diets have increased dietary plant polysaccharide fermenters (*Firmicutes*; David *et al.*, 2014). Human diet studies provide a rough outlook of how different diets could impact wild host animals' microbiomes. However, because this dataset has a likely human disease bias and uses phylum-level incidence data, diet is not a clear factor for microbiome composition. Other biases may skew nonhuman host animals' composition, such as farm and zoo environments. Therefore, the dietary and environmental diversity of mammals and birds could impact microbiome composition of non-human host animals characterized in GenBank.

Our findings further reinforce findings from an analysis by Youngblut *et al.*, where they examined the microbiome composition of animal hosts using 16S RNA sequencing, revealing four distinct subnetworks based on presence–absence data: (i) *Bacteroidetes*, (ii) *Firmicutes*, (iii) *Proteobacteria* and (iv) *Euryarchaeota* (Youngblut *et al.*, 2019). *Firmicutes* and *Proteobacteria* were found in every species, followed by *Actinobacteria* and *Bacteroidetes*, as seen in our study. *Proteobacteria* were particularly dominant in carnivores, followed by higher levels of *Bacteroidetes* in omnivores and high levels of *Firmicutes* in herbivores. In ungulates and primates, *Spirochaetes* were identified as impactful, which were also found in our ungulate data but not primate data. Actinopterygii were characterized by *Proteobacteria*, which our data also confirmed. While Youngblut *et al.* showed strong grouping by diet and taxonomic class, this was less clear with microbiome-based phylogeny. Nonetheless, the corroborating results between the Youngblut *et al.* research and our study demonstrate the promise in leveraging existing data repositories, such as GenBank.

Sequencing limitations may bias the identification of microorganisms in a given sample (Nearing *et al.*, 2021). Furthermore, the skew in GenBank towards clinical pathogens limit the ability of reference-sequence based identification of less culturable or identifiable symbionts. For example, fungal symbionts, *Euryarchaeota* and *Spirochaetes*, have been identified as having biased abundances and identification based on lack of primers and computational methods (Campanaro *et al.*, 2018; Tedersoo and Lindahl, 2016). Therefore, alongside organismal differences of diet, domestication and habitat, sequencing limitations in reference libraries (including GenBank) and primers can also play a role in determining the microbiome composition behind these results.

The results of this study provide an insight into the composition of host–microbiome data in GenBank as well as the potential for computational analysis on microbiome-based host analysis. The pathogenic focus of GenBank, while skewed when regarding microbiome analysis, can be considered in understanding microbial diseases or nonhealthy microbiomes. In addition, the database composition can be studied to understand host microbiome relationships, particularly in the context of diet and domestication as well as sequencing limitations and sampling biases. The host–microbiome pairs derived from GenBank provide unique data on often forgotten members of the microbiome, such as fungal, protistan, invertebrate and viral phyla. These phyla play smaller but valuable roles in defining the microbiome and are now being considered more in microbiome research (LaPierre *et al.*, 2019). Given that many metrics and analyses in the microbiome are often performed on solely bacteria, this study indicates the informative value of nonbacterial species in the microbiome. Another area of potential host–microbiome analysis is in non-GI isolation sources, such as skin or oral microbiomes. The framework developed for this study could therefore be extended to consider the entire group of bionts alongside the host, considering the host and its microbiome as a single evolutionary unit, known as the 'holobiont' (Huitzil *et al.*, 2018; Singh *et al.*, 2013). Finally, the results of this study provide the opportunity to consider the utility of non-traditional microbiome datasets as a complement to advance microbiome studies prospectively.

Acknowledgements

The authors thank Matthew Storer for his technical assistance with the installation and use of the original genbank-loader program. Additional technical support for this project was provided by the Center for Computation and Visualization at Brown University.

Funding

This work was funded in part by funding from the National Institutes of Health (U54GM115677).

Conflict of Interest: None declared.

References

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.

Azad,A. *et al.* (2018) HipMCL: a high-performance parallel implementation of the markov clustering algorithm for large-scale networks. *Nucleic Acids Res.*, **46**, e33.

Blacher,E. *et al.* (2017) Microbiome-modulated metabolites at the interface of host immunity. *J. Immunol.*, **198**, 572–580.

Brohee,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics.*, **7**, 488.

Campanaro,S. *et al.* (2018) Taxonomy of anaerobic digestion microbiome reveals biases associated with the applied high throughput sequencing strategies. *Sci. Rep.*, **8**, 1926.

Chabe,M. *et al.* (2017) Gut protozoa: friends or foes of the human gut microbiota? *Trends Parasitol.*, **33**, 925–934.

Chao,A. *et al.* (2014) Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.*, **84**, 45–67.
Chen,E.S. and Sarkar,I.N. (2010) MeSHing molecular sequences and clinical trials: a feasibility study. *J. Biomed. Inform.*, **43**, 442–450.
Chen,E.S. and Sarkar,I.N. (2011) Towards structuring unstructured GenBank metadata for enhancing comparative biological studies. *AMIA Jt. Summits Transl. Sci. Proc.*, **2011**, 6–10.
Chen,T. *et al.* (2017) Fiber-utilizing capacity varies in prevotella- versus bacteroides-dominated gut microbiota. *Sci. Rep.*, **7**, 2594.
Cho,I. and Blaser,M.J. (2012) The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**, 260–270.
Cho,Y.G. *et al.* (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (Oryza sativa L.). *Theor. Appl. Genet.*, **100**, 713–722.
Cryan,J.F. and O'Mahony,S.M. (2011) The microbiome-gut-brain axis: from bowel to behavior. *Neurogastroenterol. Motil.*, **23**, 187–192.
D'Argenio,V. and Salvatore,F. (2015) The role of the gut microbiome in the healthy adult status. *Clin. Chim. Acta.*, **451**, 97–102.
David,L.A. *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, **505**, 559–563.
Eckburg,P.B. *et al.* (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.
Firkins,J.L. *et al.* (2020) Extending burk dehority's perspectives on the role of ciliate protozoa in the rumen. *Front. Microbiol.*, **11**, 123.
Garmaeva,S. *et al.* (2019) Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC Biol.*, **17**, 84.
Gilbert,R.A. *et al.* (2021) Editorial: advances in the understanding of the commensal eukaryota and viruses of the herbivore gut. *Front. Microbiol.*, **12**, 619287.
Gouba,N. and Drancourt,M. (2015) Digestive tract mycobiota: a source of infection. *Med. Mal. Infect.*, **45**, 9–16.
Helmink,B.A. *et al.* (2019) The microbiome, cancer, and cancer therapy. *Nat. Med.*, **25**, 377–388.
Huitzil,S. *et al.* (2018) Modeling the role of the microbiome in evolution. *Front. Physiol.*, **9**, 1836.
Kumar,S. *et al.* (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.*, **34**, 1812–1819.
LaPierre,N. *et al.* (2019) MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. *BMC Genomics.*, **20**, 423.
Lawrence,K. and Hyde,J. (2017) Microbiome restoration diet improves digestion, cognition and physical and emotional wellbeing. *PLoS One.*, **12**, e0179017.
Ley,R.E. (2010) Obesity and the human microbiome. *Curr. Opin. Gastroenterol.*, **26**, 5–11.
Magge,A. *et al.* (2020) GeoBoost2: a natural languageprocessing pipeline for GenBank metadata enrichment for virus phylogeography. *Bioinformatics*, **36**, 5120–5121.
Magne,F. *et al.* (2020) The firmicutes/bacteroidetes ratio: a relevant marker of gut dysbiosis in obese patients? *Nutrients*, **12**, 1474.
Malla,M.A. *et al.* (2018) Exploring the human microbiome: the potential future role of Next-Generation sequencing in disease diagnosis and treatment. *Front. Immunol.*, **9**, 2868.
Mariat,D. *et al.* (2009) The firmicutes/bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.*, **9**, 123.
Martinez-Romero,E. *et al.* (2020) Plant microbiota modified by plant domestication. *Syst. Appl. Microbiol.*, **43**, 126106.
Muegge,B.D. *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.
Mukhopadhya,I. *et al.* (2019) The gut virome: the 'missing link' between gut bacteria and host immunity? *Therap. Adv. Gastroenterol.*, **12**, 1756284819836620.
Nearing,J.T. *et al.* (2021) Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, **9**, 113.
Powell,W. *et al.* (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. *Proc. Natl. Acad. Sci. U S A*, **92**, 7759–7763.
Prabhu,V.R. *et al.* (2020) Consequences of domestication on gut microbiome: a comparative study between wild gaur and domestic mithun. *Front. Microbiol.*, **11**, 133.
Reese,A.T. *et al.* (2021) Effects of domestication on the gut microbiota parallel those of human industrialization. *Elife,* **10**, e60197.
Rizzatti,G. *et al.* (2017) Proteobacteria: a common factor in human diseases. *Biomed Res. Int.*, **2017**, 9351507.
Sarkar,IN. (2010) Leveraging biomedical ontologies and annotation services to organize microbiome data from mammalian hosts. *AMIA Annu. Symp. Proc.*, **2010**, 717–721.
Sayers,E.W. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
Scotch,M. *et al.* (2011) Enhancing phylogeography by improving geographical information from GenBank. *J. Biomed. Inform.*, **44** (Suppl 1), S44–S47.
Shi,Y. *et al.* (2022) Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome*, **10**, 25.
Shin,N.R. *et al.* (2015) Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.*, **33**, 496–503.
Singh,R.K. *et al.* (2017) Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.*, **15**, 73.
Singh,Y. *et al.* (2013) Emerging importance of holobionts in evolution and in probiotics. *Gut Pathog.*, **5**, 12.
Tahsin,T. *et al.* (2016) A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. *J. Am. Med. Inform. Assoc.*, **23**, 934–941.
Tedersoo,L. and Lindahl,B. (2016) Fungal identification biases in microbiome projects. *Environ. Microbiol. Rep*, **8**, 774–779.
Vangay,P. *et al.* (2018) US immigration westernizes the human gut microbiome. *Cell*, **175**, 962–972 e910.
Wang,L.G. *et al.* (2020) Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.*, **37**, 599–603.
Weissenbacher,D. *et al.* (2015) Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics*, **31**, i348–356.
Youngblut,N.D. *et al.* (2019) Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nat. Commun.*, **10**, 2200.
Yu,G. (2020) Using ggtree to visualize data on Tree-Like structures. *Curr. Protoc. Bioinformatics.*, **69**, e96.
Zimmermann,M. *et al.* (2019) Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature*, **570**, 462–467.