BiomeSeq: A Tool for the Characterization of Animal Microbiomes from Metagenomic Data

Kelly A. Mulholland
University of Delaware
Calvin L. Keeler (ckeeler@udel.edu)
University of Delaware

Research Article

Keywords: microbiome, bacteria, BiomeSeq, microbial sequences, silico, metagenomic sequencing data

Posted Date: September 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-842545/v1

Abstract

Background

The complete characterization of a microbiome is critical in elucidating the complex ecology of the microbial composition within healthy and diseased animals. Many microbiome studies characterize only the bacterial component, for which there are several well-developed sequencing methods, bioinformatics tools and databases available. The lack of comprehensive bioinformatics workflows and databases have limited efforts to characterize the other components existing in a microbiome. BiomeSeq is a tool for the analysis of the complete animal microbiome using metagenomic sequencing data. With its comprehensive workflow and customizable parameters and microbial databases, BiomeSeq can rapidly quantify the viral, fungal, bacteriophage and bacterial components of a sample and produce informative tables for analysis.

Results

Simulated datasets were constructed, which contained known abundances of microbial sequences, and several performance metrics were analyzed, including correlation of predicted abundance with known abundance, root mean square error and rate of speed. BiomeSeq demonstrated high precision (average of 99.52%) and sensitivity (average of 93.01%). BiomeSeq was employed in detecting and quantifying the respiratory microbiome of a commercial poultry broiler flock throughout its grow-out cycle from hatching to processing and successfully processed 780 million reads. For each microbial species detected, BiomeSeq calculated the normalized abundance, percent relative abundance, and coverage as well as the diversity for each sample. Rate of speed for each step in the pipeline, precision and accuracy were calculated to examine BiomeSeq's performance using *in silico* sequencing datasets. When compared to bacterial results generated by the commonly used 16S rRNA sequencing method, BiomeSeq detected the same most abundant bacteria, including Gallibacterium, Corynebacterium and Staphylococcus, as well as several additional species.

Conclusions

BiomeSeq provides for the detection and quantification of the microbiome from next-generation metagenomic sequencing data. This tool is implemented into a user-friendly container that requires one command and generates a table containing taxonomical information for each microbe detected. It also determines normalized abundance, percent relative abundance, genome coverage and sample diversity calculations for each sample.

Introduction

Specific and unique animal microbiomes contribute to the biological function of various locations on the body including the intestinal tract, skin, vaginal tract, oral cavity, and respiratory tract [1]. Disturbances of these environments by colonization of a new bacteria, eukaryotic virus, or fungi can lead to competition, invasion and replacement. Under appropriate conditions this may result in disease. Advancements in next-generation sequencing technology enable investigations into individual components of the microbiome, thereby gaining insight into the dynamic interactions taking place [2]. Identification of microbial communities within these environments can aid in elucidating the role they play in both healthy and diseased animals.

Recent studies attempting to characterize the microbiomes of animals have focused primarily on their bacterial composition, as there are well established methodological approaches to sequence and analyze this component [3–8]. The 16S rRNA gene is commonly used to identify and compare the bacterial genera present in a given sample. Accessible bacterial databases, such as Greengenes [9] and Silva [10], in addition to well-developed bioinformatics workflows, are available to facilitate these analyses [11–13]. Internal Transcribed Spacer, or ITS, is a widely used fungal genetic marker gene. Similar to 16S rRNA, accessible fungi databases [14] and bioinformatics workflows for fungal analysis exist [12] .

Characterizing the viral component of the microbiome presents unique difficulties. Unlike the ribosomal genes of bacteria and fungi, viruses are heterogeneous in their genetic content and therefore do not have a conserved genomic region that can be sequenced and employed for taxonomic classification using the same approaches utilized for bacteria [15]. Metagenomic shotgun sequencing does not utilize gene-specific PCR amplification. As a result, this approach is not limited to detecting one specific

kingdom and has enough sensitivity to detect at the species taxonomic level. Using this approach, additional components of a microbiome can be identified. Many studies attempting to characterize microorganisms using metagenomic sequencing data rely on adapting a sequence-similarity independent assembly approach and computationally exhaustive BLAST-like database searches. This can be attributed to the limited comprehensive microbial databases that exist. Thus, this approach provides taxonomic classification of samples, but lacks the ability to accurately quantify abundance and diversity. Furthermore, many of the available computational tools require the user to possess extensive command-line knowledge and computational resources to successfully install and run the programs and their dependencies on the command line.

Herein, we present BiomeSeq, a tool for the analysis of complete animal microbiomes from metagenomic sequencing data. BiomeSeq addresses the constraints of current computational tools by providing a comprehensive workflow and corresponding microbial databases that accurately identify and quantify each major component of the microbiome. The workflow includes quality filtering and host decontamination, sequence-similarity dependent alignment to microbial reference genome databases and quantification of microbial abundance and sample diversity. BiomeSeq also analyzes the eukaryotic viral, fungal, bacteriophage and bacterial components using the same sequencing data to produce a complete analysis of the microbiome without requiring additional sequencing of the 16S rRNA or ITS genes. Utilizing shotgun metagenomic data to analyze the bacterial and fungal components can increase taxonomic resolution, permit the analysis of complete genomes instead of a conserved genomic region, and allows for a comparison of bacteria and fungi to the viral and bacteriophage components [16]. BiomeSeq was evaluated using simulated datasets designed to mimic complex microbial communities and performed with exceptional accuracy and precision. BiomeSeq was also employed to characterize the respiratory microbiome of a healthy broiler flock. The results obtained using BiomeSeq were compared to a 16S rRNA approach and BiomeSeq was able to identify 533 unique bacterial genera compared to 24 detected by 16S rRNA. In addition to characterizing all of the microbial components from the same sample, BiomeSeq is also able to discriminate at a higher taxonomic resolution. BiomeSeq is available as an open-source and user-friendly container. This versatility allows BiomeSeq to be accessible to users with varied degrees of command-line knowledge and computational resources. While BiomeSeq has been developed and evaluated on avian species, it can be used to characterize microbiomes of a variety of species.

Results

Design and Development of BiomeSeq

BiomeSeq was designed to identify microbial communities utilizing next-generation sequencing files in single- and paired-end format. Figure 1 shows an overview of the BiomeSeq workflow. In summary, the workflow begins with a quality and decontamination step where all adapter sequences, short reads and low-quality reads are first extracted from the sequencing files provided by the user. The trimmed reads are then aligned to a host reference genome specified by the user. Host DNA is extracted from the file to increase analytical efficiency and mapping accuracy [17]. The remaining reads are then aligned to BiomeSeq's eukaryotic viral, fungal, bacterial and bacteriophage genome databases containing sequences obtained from the NCBI RefSeq database. These databases are publicly available [18]. One feature that makes BiomeSeq more versatile is that it accepts custom databases provided by the user. Several additional customizable parameters can be specified by the user including: the host reference genome, mapping quality threshold, and output files (i.e. alignment files). Table 1 includes all software and parameters used by BiomeSEq. Following the alignment of the decontaminated reads to the microbial databases, BiomeSeq then calculates normalized abundance, percent relative abundance and genome coverage for each eukaryotic virus, bacteria, bacteriophage and fungi detected. It also calculates diversity of each component in the sample. BiomeSeq generates a table consisting of the NCBI RefSeq accession number, microbe name, taxonomy, number of mapped reads of the detected microbes and all calculations. For each sample processed by BiomeSeq, four tables are generated consisting of the results generated for each of the four components. In Table 2 an example of an output table for the viral component can be viewed. Similar tables are generated for bacteria, bacteriophage and fungal data. In addition to the tables, alignment files in BAM format are also generated. The BiomeSeq workflow and associated databases were implemented into a software package and a container. BiomeSeq is currently available as an open-source and user-friendly resource on Docker Hub. The self-contained environment simplifies installation and execution by eliminating the need for downloading and installing the BiomeSeq program, databases and all dependent software.

Furthermore, the BiomeSeq container allows the same customizable parameters and accepts custom databases provided by the user.

Table 1
Software tools and parameters used by BiomeSeq

Process	Tool Name	Parameters
Quality Trimming	Trim Galore	default
Host	BWA	-x -S
Decontamination	Samtools	view -bS
Microbial Database Alignment	Bowtie 2	-x -S
	Samtools	view -bSq [user input]

Table 2
Example table generated by BiomeSeq of the viral component of a commercial poultry flock at Week 6.

Ref Seq Number	Name	Taxonomy	Genome Size	Number Mapped	Norm. Abundance	Relative Abundance	Genome Coverage	Sample Diversity
NC002229	Gallid Alphaherpesvirus 2	Double Stranded; Enveloped; Herpesviridae; Mardivirus	177874	1	27	0.004%	0	0.534
NC002577	Gallid Alphaherpesvirus 3	Double Stranded; Enveloped; Herpesviridae; Mardivirus	164270	1	30	0.004%	0	
NC015396	Avian Gyrovirus	Single Stranded; Non-Enveloped; Circoviridae; Gyrovirus	2383	72	147166	22.493%	3.05	
NC001720	Fowl Aviadenovirus	Double Stranded; Non-Enveloped; Adenoviridae; Aviadenovirus	43804	4560	507049	77.498%	10.51	

Validation of BiomeSeq

BiomeSeq's performance was evaluated using simulated datasets consisting of known microorganisms and their corresponding abundances. Four simulated datasets were created to closely mimic the complex community structure of an avian respiratory microbiome. Each dataset was generated using genome sequences from 20 microorganisms that have been experimentally detected in the respiratory tract of poultry broilers (Table S1). The *Gallus gallus* red jungle fowl from one sequence was used to represent the host environment. The four simulated datasets contained an average of 24,523,032 total raw reads (Table S2) and were processed using BiomeSEq. The reads were first trimmed for quality and decontaminated of host DNA. Table S2 shows the number of reads that were extracted during each of these steps in the processing. An average of 24,522,253 remained after quality trimming of all adapter sequences, sequences less than 100 base pairs in length and sequences with a quality Phred score under 30 (Table S2). An average of 5,158,715 remained following decontamination of chicken genomic sequences. The remaining reads were then aligned to four microbial databases including bacteriophage, bacteria, fungi and avian derived virus genomes with a mapping quality threshold of 2 (available at: ref 18). One major feature of BiomeSeq that makes it so versatile, is its ability to accept custom databases provided by the user. To evaluate this feature, an avian-specific viral database was constructed to

replace BiomeSeq's default viral database (Table S5). An average of 90.8% of the reads were aligned to microbial genome sequences (Table S2). From the number of mapped reads, BiomeSeq then calculates the normalized abundance, relative abundance, genome coverage and diversity of the sample and generates a table for each of the four microbiome components. From the information provided by the BiomeSeq tables, several metrics were used to evaluate BiomeSeq's overall performance including correlation with known abundance, sensitivity, precision, rate of speed and root mean square error.

A total of twenty microbial genomes were included in the four simulated datasets and BiomeSeq was able to successfully identify each microbial element. The average percent relative abundance of each element generated by BiomeSeq was compared to the known abundances. Figure 2 shows both the known and predicted percent relative abundance of one of the simulated datasets. Of the known abundances in the simulated datasets, the most abundant fungi is *Aspergillus oryzae* with an average normalized abundance of 73.34%; the most abundant bacteria is *Escherichia coli* (10.87%); the most abundant eukaryotic virus is *Gallid Herpesvirus 2* (1.06%); and the most abundant bacteriophage is *Enterobacteriophage T4* (0.33%). The most abundant fungi detected by BiomeSeq is also *Aspergillus oryzae* with an average normalized abundance of 78.03%; the most abundant bacteria is also *Escherichia coli* (7.36%); the most abundant eukaryotic virus is *Gallid Herpesvirus 1 (*0.27%); and the most abundant bacteriophage is also *Enterobacteriophage T4* (0.09%). Pearson correlation coefficients between predicted and known abundances were calculated at the species level. Abundances of species determined by BiomeSeq were highly correlated with known abundances demonstrating an average correlation coefficient of $r = 0.997$ for all four datasets.

The precision and sensitivity of BiomeSeq was evaluated using the same four synthetic datasets. True positives, true negatives, false positives, sensitivity and precision were calculated for each microbial component (Table S3). Overall, 4,659,277 true positives, 22,397 false positives, and 350,271 false negatives were observed. Sensitivity describes the number of reads correctly aligned to the appropriate genome divided by the total number of sequences in the sample. Precision is the number of reads that were aligned to the appropriate genome divided by the total number of reads mapped to any genome. Using default parameters, BiomeSeq demonstrated exceptional accuracy, with 99.52% precision of and 93.01% sensitivity (Table S3).

The rate of speed during each step of BiomeSeq was calculated for the four simulated datasets (Fig. 3; Table S4). The speed of BiomeSeq is contingent upon the number of computational cores, the amount of available computational memory and the size of the dataset and host reference genome. The four simulated datasets were processed on a server with 98 GB RAM and 4 CPU cores. The quality step, in which adapter sequences, reads less than 100 base pairs in length and low quality reads are trimmed from the input sequencing file, was measured at an average speed of 79,977 (± 9,204) reads per second (Fig. 3; Table S4). The decontamination step had an average speed of 6,327 (± 473) reads per second (Fig. 3; Table S4). During this step, the host reference genome is indexed; the larger the host genome is the longer this step will take. The *Gallus gallus* genome (Annotation Release 104), used in this evaluation, is about 1.2 billion base pairs in length [19]. After the genome is indexed, the trimmed reads are aligned to the host reference genome and reads that map are removed from the file. Alignment of reads to microbial databases was measured at an average speed of 2,421 (± 174) reads per second (Fig. 3; Table S4). During this step, the reads remaining after decontamination, an average of 5,158,715 for the four simulated datasets, are aligned to a total of 7,227 microbial genomes with various sizes. Finally, the quantification step, in which the normalized relative abundance, percent relative abundance, genome coverage and diversity is calculated from the reads that aligned to the microbial sequences, had an average speed of 183,264 (± 31,244) reads per second (Fig. 3; Table S4).

Root mean square error (RMSE) measures the amount of error between the known abundances of each species and the abundances determined by BiomeSeq (Fig. 4). A small RMSE value indicates that the abundance determined by BiomeSeq is close to the known abundances in the simulated dataset. RMSE was calculated for each eukaryotic virus, bacteria, bacteriophage and fungi species (Fig. 4). An RMSE of < 4.70 was exhibited for all species and 17 of the 20 species exhibited an RMSE value of < 0.24. These results further indicate that BiomeSeq can accurately determine microbial abundance at the species taxonomic level.

A Longitudinal Study of the Microbial Ecology of a Healthy Broiler Flock

BiomeSeq was employed to detect and quantify eukaryotic viruses, bacteria, bacteriophage, and fungi in a healthy commercial broiler flock during the grow-out cycle from hatching to processing. Samples were collected from the respiratory tract of a healthy broiler flock weekly as the flock aged (Day 1 – Day 49). DNA and RNA were isolated and sequenced using an Illumina NGS platform. A total of 780 million reads were generated and successfully processed using BiomeSEq. These reads were sequentially

trimmed for quality, decontaminated of host DNA and aligned to each microbial genome database. The default viral genome database provided by BiomeSeq was replaced by a custom database containing avian-derived viral sequences (Table S5). For each microorganism identified, BiomeSeq calculated normalized abundance, percent relative abundance, genome coverage and sample diversity. The taxonomic and quantitative data generated by BiomeSeq was visually represented using a variety of available tools.

In total, BiomeSeq aligned 5,163 reads to avian DNA viruses and 71,936 reads to avian RNA viral sequences. A total of 9 viral species, representing 8 genera and 8 families, were identified from the avian respiratory tract during the grow-out period. Figure 5 shows a heatmap of percent normalized viral abundance at each time point during the grow-out cycle. A total of 469,937 reads were aligned to the bacterial genome database. This included 533 unique bacterial species, of which 45 had a calculated relative abundance greater than 0.5%. The 45 most abundant species detected were from 4 phyla, 7 classes, 13 orders, 26 families and 45 genera. This data is represented in a phylogenetic tree generated using the Phytools package in R (Fig. 6) [20]. A total of 504,682 reads aligned to the bacteriophage genome database. A total of 30 unique bacteriophage species from 2 orders, 5 families, and 9 genera were identified. This data is represented in a Venn diagram of the common and unique bacteriophage species detected at Week 0, Week 3 and Week 7, generated using the VennDiagram package in R (Fig. 7) [21]. A total of 1,964 reads aligned to the fungal genome database. Sixty-one unique fungal species were identified from 2 phyla, 9 classes, 20 orders, 37 families and 50 genera. This data is represented in a fungal network generated with Cytoscape in which the nodes are grouped according to class and the diameter of the inner nodes corresponds to the frequency of which that particular microbial species was detected during the grow-out cycle of the flock (Fig. 8) [22].

BiomeSeq detects the major components of a microbiome and provides the information necessary for a complete view of the microbial community's structure. To provide an example of how the taxonomic and quantitative information produced by BiomeSeq can be visually represented, a microbial network was generated using Cytoscape from a sample taken from a commercial broiler flock at 7 weeks of age (Fig. 9**)** [22]. This network contains all of the fungi, eukaryotic viruses, bacteria and bacteriophage detected by BiomeSeq, with each node diameter corresponding to percent relative abundance of the particular species detected.

A Comparison of BiomeSeq bacterial results to 16S rRNA Results

As previously discussed, 16S rRNA sequencing methods are commonly used to analyze the bacterial component of microbiome samples. To compare BiomeSeq to this method, the next generation sequencing data generated from a healthy broiler flock at week 7 was compared to 16S rRNA results. Using the same sample, metagenomic DNA Seq data and 16S rRNA data was generated. The DNA Seq data was analyzed using BiomeSeq and the16S rRNA data was analyzed using Mothur and Silva. Generally, the same bacteria were identified using both methods, although with different abundancies (Fig. 10; Table S6). BiomeSeq analysis determined that *Gallibacterium anatis* was the most abundant bacterial species (29%), followed by *Staphylococcus haemolyticus* (28%) and *Corynebacterium falsenii* (18%; **Fig. 10B**). The 16S rRNA approach determined Gallibacterium was the most abundant genra (39%), followed by Corynebacterium (23%), Lactobacillales (16%) and Staphylococcus (10%; **Fig. 10A**). BiomeSeq has greater taxonomic sensitivity and is able to identify bacteria at the species level, whereas 16S rRNA is restricted to detection at the genera level. Furthermore, BiomeSeq detected 533 unique bacteria in this sample, while 16S rRNA identified 24 genera (data not shown).

Discussion

The complete characterization of a microbiome is critical in elucidating the complex ecology of the microbial composition within healthy and diseased animals. The advancement of next generation sequencing methodologies has given rise to an increase in studies attempting to examine the microbial communities existing in a variety of animals. Readily accessible and cost-effective sequencing methodologies as well as a number of user-friendly bioinformatics analysis software and databases for 16S rRNA sequencing data provide the standard culture-independent approach for bacterial analysis [9–13]. Although 16S rRNA has provided insight into one component of the microbiome, it is limited to detecting one specific kingdom, lacks the sensitivity to discriminate between species and cannot be used for novel microbial discovery. Metagenomic shotgun sequencing does not use gene-specific PCR amplification and is therefore not restricted by primers that target specific gene sequences. As a result, it is not limited to detecting one specific kingdom and has the sensitivity to detect at the species taxonomic level. BiomeSeq is a novel

computational tool designed to characterize the complete microbiome from metagenomic sequencing data. With its comprehensive workflow and microbial reference databases, this tool can rapidly identify the eukaryotic viral, fungal, bacteriophage and bacterial components of a sample and provide an accurate quantification of abundance, genome coverage and diversity.

BiomeSeq consists of three primary steps: i) quality trimming and decontamination of host DNA; ii) alignment to four microbial reference databases; and iii) quantification of abundance, genome coverage and diversity (Fig. 1). BiomeSeq utilizes a sequence-similarity dependent approach with comprehensive microbial databases to provide taxonomic classification and quantitate abundance and diversity. This tool provides an accurate representation of abundance by considering the variability in microbial genome length and host genome length in these calculations. Comprehensive eukaryotic viral, bacterial, fungal and bacteriophage databases were constructed using complete and representative genomes obtained from the NCBI Reference Sequence Database and contained 5,693, 3,623, 1,281 and 2,212 genomes, respectively. These databases are publicly available [18]. A sequence-similarity dependent approach allows for accurate quantification; however, it is often limited by the completeness of the database used. To address this, BiomeSeq databases are updated biannually to include recently deposited microorganism sequences. Furthermore, BiomeSeq accepts custom microbial databases provided by users, thus studies are not limited to utilizing only the default databases. BiomeSeq was designed for the identification of known microorganisms, however the sequencing data accepted by this tool can also be used for *de novo* microbial discovery. Many computational tools require extensive command-line knowledge and computational resources to process sequencing samples. In an attempt to increase user accessibility, the BiomeSeq software package is implemented into an open-source and user-friendly container on DockerHub. Containers, such as this, allow the user to download and install BiomeSeq, both workflow and all databases, and dependent software on any operating system using one simple command. Furthermore, the user can process their sample with any custom parameters, using one line of code.

BiomeSeq's performance was evaluated using several metrics including correlation with known abundance, sensitivity, precision, rate of speed and root mean square error. Four simulated datasets containing known abundances of 20 microbial sequences were employed for this evaluation (Table S1). BiomeSeq was successful in identifying each of the 20 microorgansims, and the abundance calculations at the species taxonomic level determined by BiomeSeq were highly correlated with the known abundances of these species in the datasets (r = 0.997). Utilizing the default quality threshold of BiomeSeq, high precision and sensitivity were demonstrated with an average of 99.52% and 93.01%, respectively (Table S3). Rate of speed was calculated for each dataset at each step in the BiomeSeq workflow; including quality trimming, decontamination of host DNA, alignment to the four microbial databases, and quantification of relative abundance. Overall, an average total rate of speed of 271,584 (± 34,912) reads per second was observed (Fig. 3; Table S4). However, this metric is highly dependent on computational resources, as well as the size of the host reference genome and the size of the sequencing file. An RMSE of less than 0.24 was demonstrated for 17 of the species in the simulated datasets, further demonstrating that the abundance determined by BiomeSeq at the species taxonomic level corresponds to the known values (Fig. 4). Overall, BiomeSeq performed with exceptional speed, accuracy and sensitivity.

Biomeseq was employed to detect and quantify the respiratory microbiome of a healthy commercial poultry broiler flock at weekly intervals from hatching to processing. For each component of the respiratory microbiome of this flock, abundance was calculated and population shifts were examined at each time point. A total of 11 eukaryotic viral species, 45 bacterial species, 31 bacteriophage species, and 61 fungal species were identified in this flock. The taxonomic and quantitative tables generated by BiomeSeq can be input into several programs to create visual representations of the data. Heatmaps, phylogenetic trees, Venn diagrams, and microbial networks are examples of visualizations that can be easily generated to assist interpretation of the results (Figs. 5–9).

The commercial broiler flock utilized in this study was vaccinated *in ovo* with a live Marek's disease virus vaccine (SB-1) and a live recombinant herpesvirus of turkeys (HVT) vaccine expressing Newcastle disease virus genes. The presence of herpesviruses and coronaviruses in the respiratory tract is consistent with vaccination with these two live vaccines, coupled with the expected presence of these avian viruses in the environment. The presence of the Myoviridae family of bacteriophage correlated with the presence of Gallibacterium, an abundant commensal avian bacterial species (data not shown). Interactions between bacteriophage and bacteria are known to have a significant impact on host health (24). Basidiomycota was highly abundant in

this flock, however further studies are needed to determine the relevance of this fungal species in the respiratory tract of avian species. The bacterial diversity of the flock was complex at the time of processing, containing significant amounts of Pasteurellaceae, Corynebacteriaceae, Staphylococcaceae and Enterobacteriaceae.

Using one sample from this study, bacterial results generated by BiomeSeq were compared to results generated by 16S rRNA sequencing methods. The most abundant bacteria species were commonly? identified using both methods (Fig. 11; Table S6). However, BiomeSeq identified 533 unique bacterial species, 45 with a relative abundance of greater than 0.5%, while 16S rRNA detected only 24 genera. Furthermore, BiomeSeq has greater taxonomic sensitivity and is able to identify bacteria at the species level, whereas 16S rRNA sequence analysis is restricted to detection at the genera level. Moreover, 16S rRNA sequencing methodology can only be employed for taxonomic classification of the bacterial component, leaving the identity of the remaining components of the microbiome unknown. BiomeSeq is able to characterize all major components of a microbiome with high taxonomic sensitivity and it accurately quantifies abundance. Moreover, unlike 16S rRNA sequencing data, metagenomic shotgun sequencing data processed by BiomeSeq can be further used in sequence-independent approaches for *de novo* microbial discovery.

Conclusions

BiomeSeq was developed for the analysis of complete animal microbiomes using metagenomic sequencing data. With its comprehensive workflow, customizable parameters and associated microbial databases, BiomeSeq can rapidly identify the major components of a microbiome from a biological sample and determine normalized abundance, percent relative abundance, genome coverage and sample diversity. While many existing tools focus on characterizing one microorganism, BiomeSeq provides a complete view of microbial ecology and diversity in a sample. The performance of this tool was evaluated using both simulated and clinical datasets and accurate and precise abundance estimates were demonstrated. BiomeSeq is available as an open-source and user-friendly container, allowing users to easily download, install and use the program with a few simple commands. The versatility of BiomeSeq, such as customizable parameters and accepting custom databases, allow this tool to facilitate a variety of unique investigations.

Methods

Cookie PolicyCookie Policy Research Square My Dashboard MY ARTICLE ADMIN Workflows Draft AUTHOR'S VIEW Peer Review Timeline Preprint Preview TOOLS & SERVICES In ReviewBMC Genomics 736aaa7a-8585-4f74-b34a-9ebd636dbb2a | r1 RESEARCH ARTICLE Epigenetics & Genomics BiomeSeq: A Tool for the Characterization of Animal Microbiomes from Metagenomic Data Kelly A. Mulholland, Calvin L. Keeler Abstract Background The complete characterization of a microbiome is critical in elucidating the complex ecology of the microbial composition within healthy and diseased animals. Many microbiome studies characterize only the bacterial component, for which there are several well-developed sequencing methods, bioinformatics tools and databases available. The lack of comprehensive bioinformatics workflows and databases have limited efforts to characterize the other components existing in a microbiome. BiomeSeq is a tool for the analysis of the complete animal microbiome using metagenomic sequencing data. With its comprehensive workflow and customizable parameters and microbial databases, BiomeSeq can rapidly quantify the viral, fungal, bacteriophage and bacterial components of a sample and produce informative tables for analysis. Results Simulated datasets were constructed, which contained known abundances of microbial sequences, and several performance metrics were analyzed, including correlation of predicted abundance with known abundance, root mean square error and rate of speed. BiomeSeq demonstrated high precision (average of 99.52%) and sensitivity (average of 93.01%). BiomeSeq was employed in detecting and quantifying the respiratory microbiome of a commercial poultry broiler flock throughout its grow-out cycle from hatching to processing and successfully processed 780 million reads. For each microbial species detected, BiomeSeq calculated the normalized abundance, percent relative abundance, and coverage as well as the diversity for each sample. Rate of speed for each step in the pipeline, precision and accuracy were calculated to examine BiomeSeq's performance using in silico sequencing datasets. When compared to bacterial results generated by the commonly used 16S rRNA sequencing method, BiomeSeq detected the same most abundant bacteria, including Gallibacterium, Corynebacterium and Staphylococcus, as well as several additional species. Conclusions BiomeSeq provides for the detection and quantification of the microbiome from next-generation metagenomic sequencing data. This tool is implemented into a user-friendly container that requires one command and

generates a table containing taxonomical information for each microbe detected. It also determines normalized abundance, percent relative abundance, genome coverage and sample diversity calculations for each sample. KEYWORDS microbiome, bacteria, BiomeSeq, microbial sequences, silico, metagenomic sequencing data Fulltext source File 736aaa7a-8585-4f74-b34a-9ebd636dbb2a_-1_enriched.html User ACDC Connector Updated at September 7 2021, 2:59:11 AM GMT+5:30 Edit Fulltext Introduction Results Discussion Conclusions Methods

BiomeSeq is currently available as an open-access and user-friendly tool on Docker Hub. As the docker container is self-contained, it simplifies installation and execution by eliminating the need for downloading and installing dependent software and requires only one command. BiomeSeq is customizable and allows the user to adjust parameters similar to a command-line tool. Table 1 includes all software and parameters used in BiomeSeq.

BiomeSeq accepts both single- and paired-end reads in fastq format generated by DNA Seq or RNA Seq methods. Along with the fastq file, the user may customize a number of parameters including: the host genome that the sample was derived from, custom databases provided by the user, mapping quality threshold and output file types. Figure 1 shows an overview of the BiomeSeq workflow, which consists of three primary steps: i) quality trimming and decontamination of host DNA; ii) alignment to four microbial reference databases; and iii) quantification of abundance, genome coverage and diversity. BiomeSeq generates a table consisting of NCBI RefSeq accession number, microbe name, taxonomic information, number of mapped reads, normalized abundance, percent relative abundance, genome coverage for each eukaryotic virus, bacteria, bacteriophage and fungi detected, as well as the alpha diversity of the sample. Table 2 is an example of an output table generated for a eukaryotic virus anlaysis. Similar tables are produced for bacteria, bacteriophage and fungal data. In addition to the tables, the option to output alignment files in BAM format is also available. Visualizations of these results can be easily generated using several different packages in R.

Quality Trimming and Host Decontamination

The BiomeSeq workflow begins with a quality trimming step in which individual fastq sequence files are analyzed for per-base sequence quality, per-sequence quality, sequence length distribution and duplicate sequences (Fig. 1). Reads with a quality phred score below 30, reads under 100 base pairs in length and adapter sequences are removed from the file. This step is conducted using Trim-Galore [23]. The next step in the workflow decontaminates the file of host DNA. In this step, the trimmed reads are aligned to the user-specified host reference genome using BWA, and only reads that do not align to the host genome are extracted and analyzed further (Fig. 1) [24].

Microbial Database Alignment

The trimmed and decontaminated sequencing reads are aligned to a eukaryotic viral genome database, a bacterial database, a fungal database and a bacteriophage database using the Bowtie 2 alignment algorithm (Fig. 1) [25]. The mapping quality threshold default is 20, however this parameter may be customized by the user. The eukaryotic viral genome database currently includes 5,693 complete and representative viral sequences obtained from the National Center for Biotechnology Information (NCBI) Reference Sequence Database [26]. Bacterial, fungal and bacteriophage databases were constructed using a similar approach and contain 3,623, 1,281 and 2,212 genomes, respectively [26]. Each microbial database and corresponding aligner index files are publicly available [18]. Each of the four microbial databases are continuously updated to include novel and recently discovered sequences. These databases are the default option for BiomeSEq. However, as an additional feature, BiomeSeq also accepts custom microbial databases provided by the user, as was demonstrated in this study.

Quantification and Output

A sequence similarity-dependent approach for detecting microorganisms contributes to the rapid detection of known viruses while also allowing for the quantification of biodiversity, which similarity-independent approaches lack [27, 28]. To calculate microbial abundance, BiomeSeq uses an adaptation to the equation presented by Moustafa and colleagues in 2017 to quantify viral abundance [29]:

$$Microbial\ Abundance = \frac{2\ x\ \frac{number\ of\ reads\ mapped\ to\ microbe\ sequence}{microbe\ sequence\ size}}{\frac{number\ of\ reads\ mapped\ to\ host\ genome}{host\ genome\ size}}\ x\ 10^5$$

Percent relative abundance is quantified using the following equation:

$$Percent\ Relative\ Abundance = \frac{microbial\ abundance}{total\ microbial\ abundance} x\ 100$$

Genome coverage is approximated using the following equation:

$$Genome\ Coverage = \frac{(number\ of\ reads\ mapped\ to\ microbe\ \ x\ \ read\ length)}{microbe\ reference\ genome\ size}$$

Alpha diversity for each sample is calculated using the Shannon Diversity Index, a commonly used equation for calculating species diversity in a microbiome as it accounts for both species abundance and evenness within the sample [30, 31].

Performance Metrics

Simulated data was utilized to assess several metrics of BiomeSeq's performance capabilities including correlation with known abundance, sensitivity, precision, rate of speed and root mean square error. Four datasets were generated to closely mimic the complexity of data obtained from real microbiomes consisting of bacteria, eukaryotic viruses, bacteriophage and fungi genomes as well as host DNA sequences. The datasets contain sequences from 20 microorganisms commonly found in the respiratory microbiome of broiler chickens, including 10 eukaryotic viruses, 4 bacteriophage, 5 bacteria and 1 fungi (Table S1). We included one chicken sequence to represent the host environment (NC_006088.5). ART was used to simulate reads generated using next-generation sequencing technology [32]. Single-end reads with a length of 100, fold coverage of 10X and masking cutoff frequency of 1 in 100 were simulated based on an error and quality profile of the HiSeq 2500 Illumina sequencing platform. The number of reads simulated ranged from 24,522,223 to 24,523,708, with an average read count of 24,523,065.

The four simulated datasets were processed using BiomeSeq with the following parameters: -g chicken.fasta -d avian_virus -q 20. One major feature of BiomeSeq is its ability to accept custom databases provided by the user. To evaluate this feature, an avian-specific viral database was constructed to replace BiomeSeq's default viral database (Table S5). The avian DNA viral genomes include 48 viral elements from 9 unique families and the avian RNA viral genomes include 63 viral elements from 13 families. The avian DNA and RNA viral database is arranged by the classification of their viral structure and genome organization. DNA viruses are organized hierarchically by whether the virus is double- or single-stranded and whether the virus is enveloped or non-enveloped. RNA viruses are organized hierarchically by whether the virus is double- or single-stranded, negative or positive sense, segmented or non-segmented and whether the virus is enveloped or non-enveloped. This database is publicly available [18]. Abundance was calculated on the species level and several metrics were assessed based on the calculations determined by BiomeSeq including correlation with known abundances which was calculated using Pearson's correlation coefficient. Rate of speed was calculated as the number of reads per second at each step of the BiomeSeq process on a server with 98 GB RAM and 4 CPU cores. Sensitivity and precision were calculated based on the following equations:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

True positives are the number of reads that BiomeSeq aligned to the genomes in the databases; false positives are the number of reads that were aligned to genomes not included in the databases; and false negatives are the number of reads that were not aligned. Root mean square error was calculated to compare the abundance calculations of BiomeSeq to the known abundance using the following equation:

$$RMSE = \sqrt{\frac{\sum(Biomeseq\ Abundance - Known\ Abundance)^2}{number\ samples}}$$

A Longitudinal Study of the Microbial Ecology of a Healthy Broiler Flock

Tracheal swabs were collected at hatching and at weekly intervals through processing at day 49 (8 samples) from an antibiotic-free commercial broiler flock. Both DNA and RNA were isolated and sequencing was performed for each of the eight time points using the Illumina HiSeq platform producing 1 X 100 single-end reads. Each of the resulting 16 samples were processed using BiomeSeq with the following parameters: -g chicken -d avian_virus -q 40. The previously described avian viral reference database was utilized in this study (Table S5). Normalized abundance, relative abundance, genome coverage and sample diversity was calculated for each of the microbial components. Visual representations of the results generated by BiomeSeq were generated using several R packages, including heatmaps, phylogenetic trees, Venn diagrams and microbial networks [20–22].

Comparison of BiomeSeq Bacterial Results to 16S rRNA Results

For comparison of BiomeSeq results to bacterial results generated using 16S rRNA sequencing methodology, the V4 hypervariable region of the bacterial 16S rRNA gene was extracted and amplified using PCR with primers 515F (‘5- GTGCCAGCMGCCGCGGTAA-3’) and 806R (‘5-GGACTACHVGGGTWTCTAAT-3’), as previously described [7, 33]. The amplicons were sequenced at the University of Minnesota Genomics Center (Minneapolis, MN) using an Illumina MiSeq 600 cycle v3 kit. Each sample was assessed for quality and assembled into contigs using PEAR’s default parameters, with the modification that the quality score threshold was set to 30. Samples were further filtered and analyzed using mothur version 1.35.1 [13] and MiSeq SOP [34]. OTUs were generated using 97% sequence similarity. Mothur’s implementation of the SILVA database (v123) was used for classification of OTUs, and relative abundance was calculated. The results generated using 16S rRNA sequencing methodology were compared to results generated by BiomeSeq.

Preview BiomeSeq is currently available as an open-access and user-friendly tool on Docker Hub. As the docker container is self-contained, it simplifies installation and execution by eliminating the need for downloading and installing dependent software and requires only one command. BiomeSeq is customizable and allows the user to adjust parameters similar to a command-line tool. Table 1 includes all software and parameters used in BiomeSeq. BiomeSeq accepts both single- and paired-end reads in fastq format generated by DNA Seq or RNA Seq methods. Along with the fastq file, the user may customize a number of parameters including: the host genome that the sample was derived from, custom databases provided by the user, mapping quality threshold and output file types. Figure 1 shows an overview of the BiomeSeq workflow, which consists of three primary steps: i) quality trimming and decontamination of host DNA; ii) alignment to four microbial reference databases; and iii) quantification of abundance, genome coverage and diversity. BiomeSeq generates a table consisting of NCBI RefSeq accession number, microbe name, taxonomic information, number of mapped reads, normalized abundance, percent relative abundance, genome coverage for each eukaryotic virus, bacteria, bacteriophage and fungi detected, as well as the alpha diversity of the sample. Table 2 is an example of an output table generated for a eukaryotic virus anlaysis. Similar tables are produced for bacteria, bacteriophage and fungal data. In addition to the tables, the option to output alignment files in BAM format is also available. Visualizations of these results can be easily generated using several different packages in R. Quality Trimming and Host Decontamination The BiomeSeq workflow begins with a quality trimming step in which individual fastq sequence files are analyzed for per-base sequence quality, per-sequence quality, sequence length distribution and duplicate sequences (Fig. 1). Reads with a quality phred score below 30, reads under 100 base pairs in length and adapter sequences are removed from the file. This step is conducted using Trim-Galore [23]. The next step in the workflow decontaminates the file of host DNA. In this step, the trimmed reads are aligned to the user-specified host reference genome using BWA, and only reads that do not align to the host genome are extracted and analyzed further (Fig. 1) [24]. Microbial Database Alignment The trimmed and decontaminated sequencing reads are aligned to a eukaryotic viral genome database, a bacterial database, a fungal database and a bacteriophage database using the Bowtie 2 alignment

algorithm (Fig. 1) [25]. The mapping quality threshold default is 20, however this parameter may be customized by the user. The eukaryotic viral genome database currently includes 5,693 complete and representative viral sequences obtained from the National Center for Biotechnology Information (NCBI) Reference Sequence Database [26]. Bacterial, fungal and bacteriophage databases were constructed using a similar approach and contain 3,623, 1,281 and 2,212 genomes, respectively [26]. Each microbial database and corresponding aligner index files are publicly available [18]. Each of the four microbial databases are continuously updated to include novel and recently discovered sequences. These databases are the default option for BiomeSEq. However, as an additional feature, BiomeSeq also accepts custom microbial databases provided by the user, as was demonstrated in this study. Quantification and Output A sequence similarity-dependent approach for detecting microorganisms contributes to the rapid detection of known viruses while also allowing for the quantification of biodiversity, which similarity-independent approaches lack [27, 28]. To calculate microbial abundance, BiomeSeq uses an adaptation to the equation presented by Moustafa and colleagues in 2017 to quantify viral abundance [29]: Percent relative abundance is quantified using the following equation: Genome coverage is approximated using the following equation: Alpha diversity for each sample is calculated using the Shannon Diversity Index, a commonly used equation for calculating species diversity in a microbiome as it accounts for both species abundance and evenness within the sample 30, 31. Performance Metrics Simulated data was utilized to assess several metrics of BiomeSeq's performance capabilities including correlation with known abundance, sensitivity, precision, rate of speed and root mean square error. Four datasets were generated to closely mimic the complexity of data obtained from real microbiomes consisting of bacteria, eukaryotic viruses, bacteriophage and fungi genomes as well as host DNA sequences. The datasets contain sequences from 20 microorganisms commonly found in the respiratory microbiome of broiler chickens, including 10 eukaryotic viruses, 4 bacteriophage, 5 bacteria and 1 fungi (Table S1). We included one chicken sequence to represent the host environment (NC_006088.5). ART was used to simulate reads generated using next-generation sequencing technology 32 . Single-end reads with a length of 100, fold coverage of 10X and masking cutoff frequency of 1 in 100 were simulated based on an error and quality profile of the HiSeq 2500 Illumina sequencing platform. The number of reads simulated ranged from 24,522,223 to 24,523,708, with an average read count of 24,523,065. The four simulated datasets were processed using BiomeSeq with the following parameters: -g chicken.fasta -d avian_virus -q 20. One major feature of BiomeSeq is its ability to accept custom databases provided by the user. To evaluate this feature, an avian-specific viral database was constructed to replace BiomeSeq's default viral database (Table S5). The avian DNA viral genomes include 48 viral elements from 9 unique families and the avian RNA viral genomes include 63 viral elements from 13 families. The avian DNA and RNA viral database is arranged by the classification of their viral structure and genome organization. DNA viruses are organized hierarchically by whether the virus is double- or single-stranded and whether the virus is enveloped or non-enveloped. RNA viruses are organized hierarchically by whether the virus is double- or single-stranded, negative or positive sense, segmented or non-segmented and whether the virus is enveloped or non-enveloped. This database is publicly available 18 . Abundance was calculated on the species level and several metrics were assessed based on the calculations determined by BiomeSeq including correlation with known abundances which was calculated using Pearson's correlation coefficient. Rate of speed was calculated as the number of reads per second at each step of the BiomeSeq process on a server with 98 GB RAM and 4 CPU cores. Sensitivity and precision were calculated based on the following equations: True positives are the number of reads that BiomeSeq aligned to the genomes in the databases; false positives are the number of reads that were aligned to genomes not included in the databases; and false negatives are the number of reads that were not aligned. Root mean square error was calculated to compare the abundance calculations of BiomeSeq to the known abundance using the following equation: A Longitudinal Study of the Microbial Ecology of a Healthy Broiler Flock Tracheal swabs were collected at hatching and at weekly intervals through processing at day 49 (8 samples) from an antibiotic-free commercial broiler flock. Both DNA and RNA were isolated and sequencing was performed for each of the eight time points using the Illumina HiSeq platform producing 1 X 100 single-end reads. Each of the resulting 16 samples were processed using BiomeSeq with the following parameters: -g chicken -d avian_virus -q 40. The previously described avian viral reference database was utilized in this study (Table S5). Normalized abundance, relative abundance, genome coverage and sample diversity was calculated for each of the microbial components. Visual representations of the results generated by BiomeSeq were generated using several R packages, including heatmaps, phylogenetic trees, Venn diagrams and microbial networks 20– 22. Comparison of BiomeSeq Bacterial Results to 16S rRNA Results For comparison of BiomeSeq results to bacterial results generated using 16S rRNA sequencing methodology, the V4 hypervariable region of the bacterial 16S rRNA gene was extracted and amplified using PCR with primers 515F ('5- GTGCCAGCMGCCGCGGTAA-3') and 806R ('5-GGACTACHVGGGTWTCTAAT-3'), as previously described 7, 33. The amplicons were sequenced at the University of Minnesota Genomics Center (Minneapolis, MN) using an Illumina MiSeq 600 cycle v3 kit. Each sample was assessed for quality and assembled into contigs using PEAR's default parameters, with the modification

that the quality score threshold was set to 30. Samples were further filtered and analyzed using mothur version 1.35.1 13 and MiSeq SOP 34 . OTUs were generated using 97% sequence similarity. Mothur's implementation of the SILVA database (v123) was used for classification of OTUs, and relative abundance was calculated. The results generated using 16S rRNA sequencing methodology were compared to results generated by BiomeSeq. Abbreviations Declarations References Unsectioned Paragraphs Competing Interests Add associated publications Manuscript Files Add figures Add Funders Supplementary Files Reference Files PRODUCTION WORKFLOW Stage: Production/InProgress Previous step notes: ACDC Annotations: Annotation for presence of Bio-security agents keyword (Escherichia coli, vaccination, vaccines) Annotation for identifying "Coronavirus" related terms (coronaviruses) - ACDC C Owner: Pradeep Gaikwad Notes: Transition: Complete Production is complete and QC can begin. Relinquish Remove your claim to the production stage and move back to "ready". Time Spent (minutes): StoA - In Review Journal QC Passed EDITOR ACTIONS Tags Select... Research Square Research Square lets you share your work early, gain feedback from the community, and start making changes to your manuscript prior to peer review in a journal. As a division of Research Square Company, we're committed to making research communication faster, fairer, and more useful. We do this by developing innovative software and high quality services for the global research community. Our growing team is made up of researchers and industry professionals working together to solve the most critical problems facing scientific publishing. Twitter logo Facebook logo YouTube logo Vimeo logo Instagram logo LinkedIn logo Also discoverable on ResearcherApp logo PLATFORM About Our Team In Review Editorial Policies Advisory Board Contact Us RESOURCES Author Services Blog Accessibility API Access RSS feed COMPANY About Us Careers Partner With Us Responsibility Press GET UPDATES First Name Last Name Email SUBSCRIBE © Research Square 2021 | ISSN 2693-5015 (online) Privacy Policy Terms of Service Do Not Sell My Personal Information Loading... | Research Square

Abbreviations

16S rRNA: 16S ribosomal RNA

NCBI: National Center for Biotechnology Information

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

RMSE: root mean square error

BLAST: Basic Local Alignment Search Tool

Declarations

Ethics Statement

Not applicable.

Consent for publication

All authors have consented to publication

Availability of data and material

The BiomeSeq Docker container is available at http://dockerhub.com.

BiomeSeq custom databases are available at https://de.cyverse.org.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported by Agriculture and Food Research Initiative Competitive Grant #2015-68004-23131 from the USDA National Institute of Food and Agriculture. Computational infrastructure support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible through funding from Delaware INBRE (NIH P20 GM103446) and the Delaware Biotechnology Institute.

Authors' contributions

KAM is the primary author of this manuscript. KAM designed and developed the bioinformatics workflow and constructed the avian-derived viral genome database, the eukaryotic viral database, the fungal database, the bacteriophage database and the bacterial database. KAM wrote all programs for microbial calculations and programs to generate visual representations of microbial data. CLK is the corresponding author of this work. CLK contributed to the design of the work, the acquisition of samples, the analysis and interpretation of the data, and revised and edited the manuscript.

Acknowledgements

We thank Monique Robinson, Sharon Keeler, Hong Li and Daniel Bautista for their contributions in collecting and processing experimental samples and Shawn Polson for his insightful comments and suggestions for this manuscript.

References

1. The NIH Human Microbiome Project. Genome research, 2009. **19**(12): p. 2317-2323.
2. Barzon, L., et al., *Applications of next-generation sequencing technologies to diagnostic virology.* Int J Mol Sci, 2011. **12**(11): p. 7861-84.
3. Bond, S.L., et al., *Upper and lower respiratory tract microbiota in horses: bacterial communities associated with health and mild asthma (inflammatory airway disease) and effects of dexamethasone.* BMC Microbiol, 2017. **17**(1): p. 184.
4. De Boeck, C., et al., *Longitudinal monitoring for respiratory pathogens in broiler chickens reveals co-infection of Chlamydia psittaci and Ornithobacterium rhinotracheale.* J Med Microbiol, 2015. **64**(Pt 5): p. 565-74.
5. Gaeta N, L.S., Teixeira A, Ganda E, Oikonomou G, Gregory L, Bichalho R., *Deciphering upper respiratory tract microbiota complexity in healthy calves and calves that develop respiratory disease using shotgun metagenomics.* J Dairy Sci. , 2017. **100**: p. 1445-1458.
6. Glendinning, L., G. McLachlan, and L. Vervelde, *Age-related differences in the respiratory microbiota of chickens.* PLoS One, 2017. **12**(11): p. e0188455.
7. Johnson TJ, Y.B., Noll S, Cardona C, Evans NP, Karnezos P, Ngunjiri JM, Abundo MC, Lee C-W, *A consistent and predictable commercial broiler chicken bacterial microbiota in antibiotic-free production displays strong correlations with performance.* Appl. Environ. Micro., 2018. **84**: p. e00362-18.
8. Shabbir, M.Z., et al., *Microbial communities present in the lower respiratory tract of clinically healthy birds in Pakistan.* Poult Sci, 2015. **94**(4): p. 612-20.
9. De Santis T, H.P., Larsen N, Rojas M, Brodie E, Keller K, Huber T, Dalevi D, Hu P, Andersen G., *Greengenes, a chimera-checked 16S rRNA gene.* 2016.
10. Quast C, P.E., Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner F., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucl Acids Res., 2013. **41**: p. 590-596.
11. Meyer F, P.D., D'Souza M, Olson R, Glass E, Kubal M, Paczian T, Rodriguqz A, Stevens R, Wilke A, Wilkening J, Edwards R., *The metagenomics RAST server- a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**: p. 386.
12. Caporaso J, K.J., Stombaugh J, Bittinger K, Bushman F, Costello E, Fierer N, Peña A, Goodrich J, Gordon J, Huttley G, Kelley ST, Knights D, Koenig JE, Ley R, Lozupone C, McDonald D, Muegge B, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh P, Walters W, Widmann J, Yatsunenko T, Zaneveld J, Knight R., *Qiime allows analysis of high-throughout community sequencing data.* Nature Methods, 2010. **7**: p. 335-336.

13. Schloss P, W.S., Ryabin T, Hall J, Hartman M, Hollister E, Lesniewski R, Oakley B, Parks D, Robinson C, Sahl J, Stres B, Thallinger G, Van Horn D, Weber C. , *Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl Enviro Microbiol, 2009. **75**: p. 7537-7541.
14. Kõljalg U, N.R., Abarenkov K, Tedersoo L, Taylor A, Bahram M, Bates S, Bruns T, Bengtsson-Palme J, Callaghan T, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith G, Hartmann M, Kirk P, Kohout P, Larsson E, Lindahl B, Lücking R, Martín M, Matheny P, Nguyen N, Niskanen T, Oja J, Peay K, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott J, Senés C, Smith M, Suija A, Taylor D, Telleria M, Weiss M, Larsson K., *Towards a unified paradigm for sequence-based identification of fungi.* Mol Ecol., 2013. **22**: p. 5271-5277.
15. Zhu, J., et al., *Virus-specific CD8+ T cells accumulate near sensory nerve endings in genital skin during subclinical HSV-2 reactivation.* (0022-1007 (Print)).
16. Jovel, J., et al., *Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics.* Front Microbiol, 2016. **7**: p. 459.
17. Daly G, L.R., Rowe W, Stubbs S, Wilkinson M, Ramirez-Gonzalez R, Mario C, Bernal W, Heeney J. , *Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data.* PLoS One, 2015. **10**(6).
18. Mulholland, K.A. *BiomeSeq Microbial Databases*. Avian Genomics 2019; Available from: https://sites.udel.edu/aviangenomics/.
19. Hillier, L.W., et al., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.
20. L., R., *Phytools: An R package for phylogenetic comparative biology (and other things).* Methods Ecol. Evol., 2012. **3**: p. 217-223.
21. Chen, H. and P.C. Boutros, *VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.* BMC Bioinformatics, 2011. **12**(1): p. 35.
22. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* (1088-9051 (Print)).
23. M., M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet Journal, 2011. **17**: p. 10-12.
24. Li H., D.R., *Fast and accurate long-read alignment with Burrows-Wheeler Transform.* Bioinformatics, 2009. **EPub**.
25. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
26. O'Leary NA, W.M., Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. , *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res., 2016. **4**: p. 733-745.
27. Herath, D., et al., *Assessing Species Diversity Using Metavirome Data: Methods and Challenges.* Comput Struct Biotechnol J, 2017. **15**: p. 447-455.
28. Rose, R., et al., *Challenges in the analysis of viral metagenomes.* Virus Evol, 2016. **2**(2): p. vew022.
29. Moustafa, A., et al., *The blood DNA virome in 8,000 humans.* PLoS Pathog, 2017. **13**(3): p. e1006292.
30. Lemos, L.N., et al., *Rethinking microbial diversity analysis in the high throughput sequencing era.* Journal of Microbiological Methods, 2011. **86**(1): p. 42-51.
31. Ludwig, J. and J. Reynolds, *Statistical Ecology*, ed. Wiley. 1988, New York.
32. Huang, W., et al., *ART: a next-generation sequencing read simulator.* Bioinformatics, 2012. **28**(4): p. 593-4.
33. Gohl DM, V.P., Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D, Beckman KB., *Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies.* Nat Biotechnol, 2016. **34**: p. 942-949.
34. Kozich J, W.S., Baxter N, Highlander S, Schloss P., *Development of a dual-index se-quencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform.* Appl. Environ. Microbiol., 2013. **79**: p. 5112-

5120.

Figures

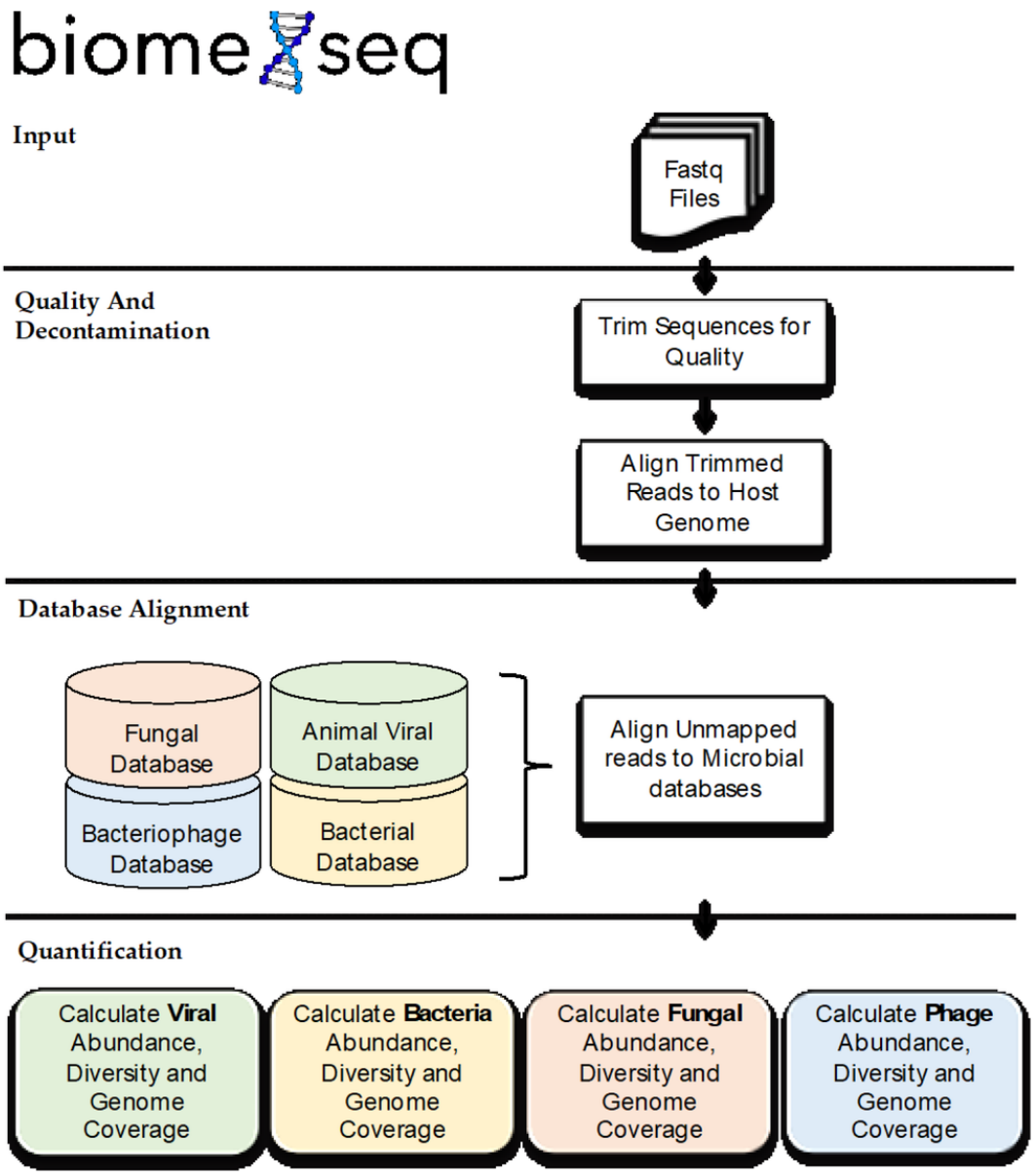

Figure 1

BiomeSeq Workflow

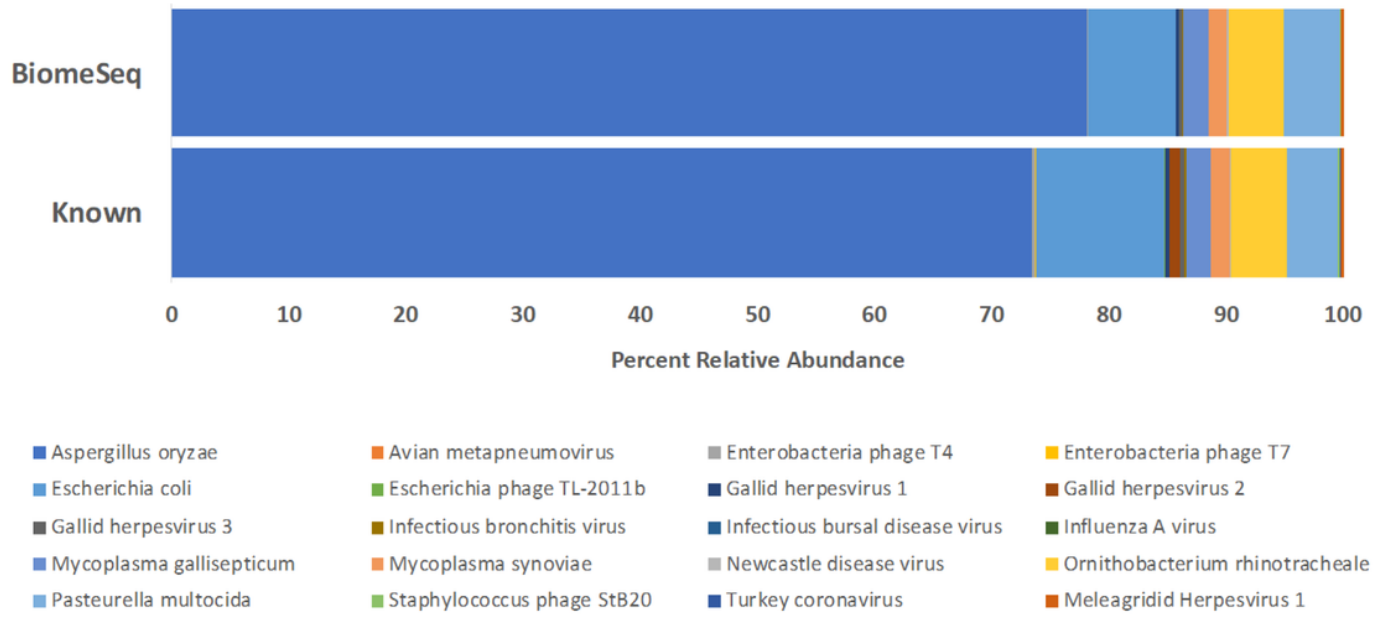

Figure 2

Percent relative abundance of microorganisms detected by BiomeSeq and known values from one simulated dataset.

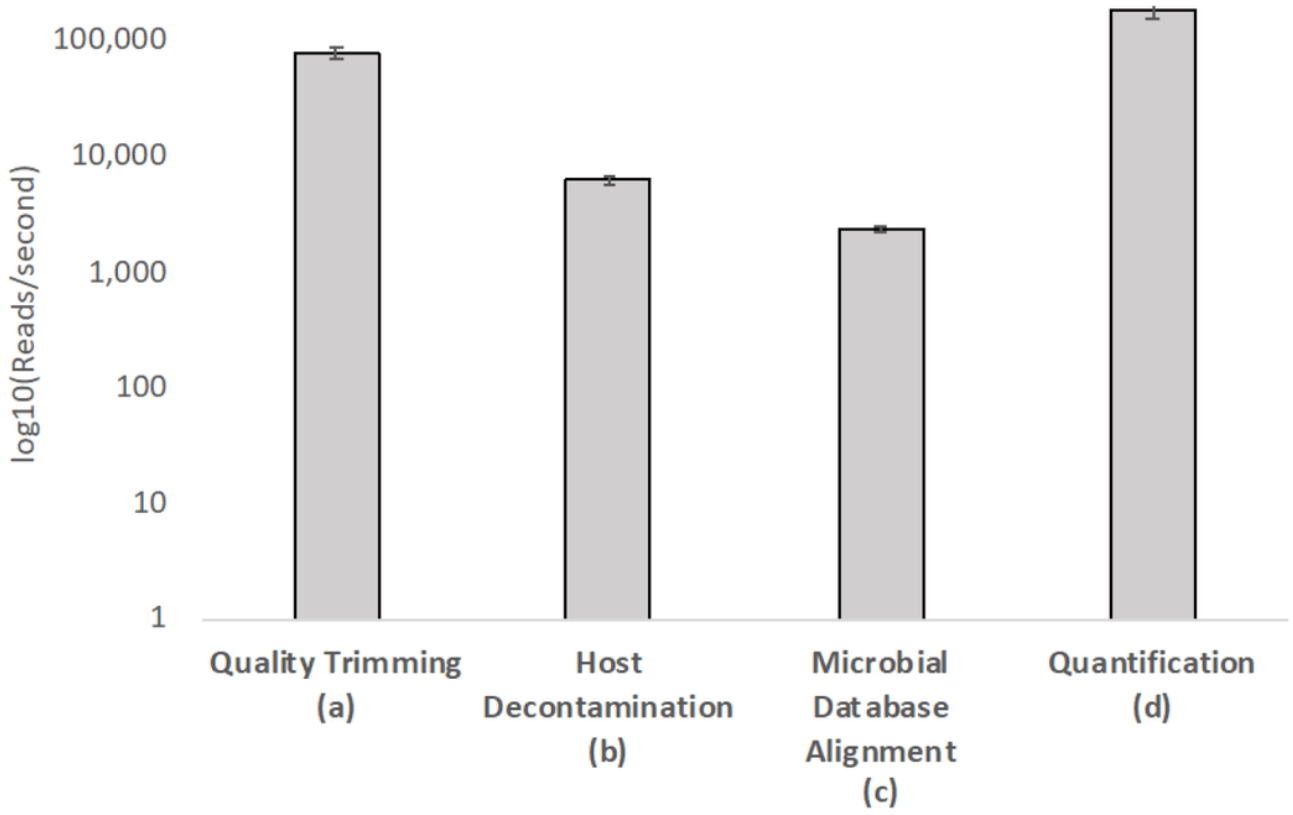

Figure 3

Average rate of speed at different steps in BiomeSeq processing for four simulated datasets. a) Adapter sequences, reads shorter than 100 base pairs in length and reads with a quality Phred score of less than 30 from the sequencing file. b) Host reference genome is indexed, and reads are aligned to the host reference genome to extract host DNA. c) Reads are aligned to four microbial databases including eukaryotic viruses, fungi, bacteria and bacteriophage. d) Normalized abundance, percent relative abundance, genome coverage and diversity are calculated from the reads that align to microbial sequences.

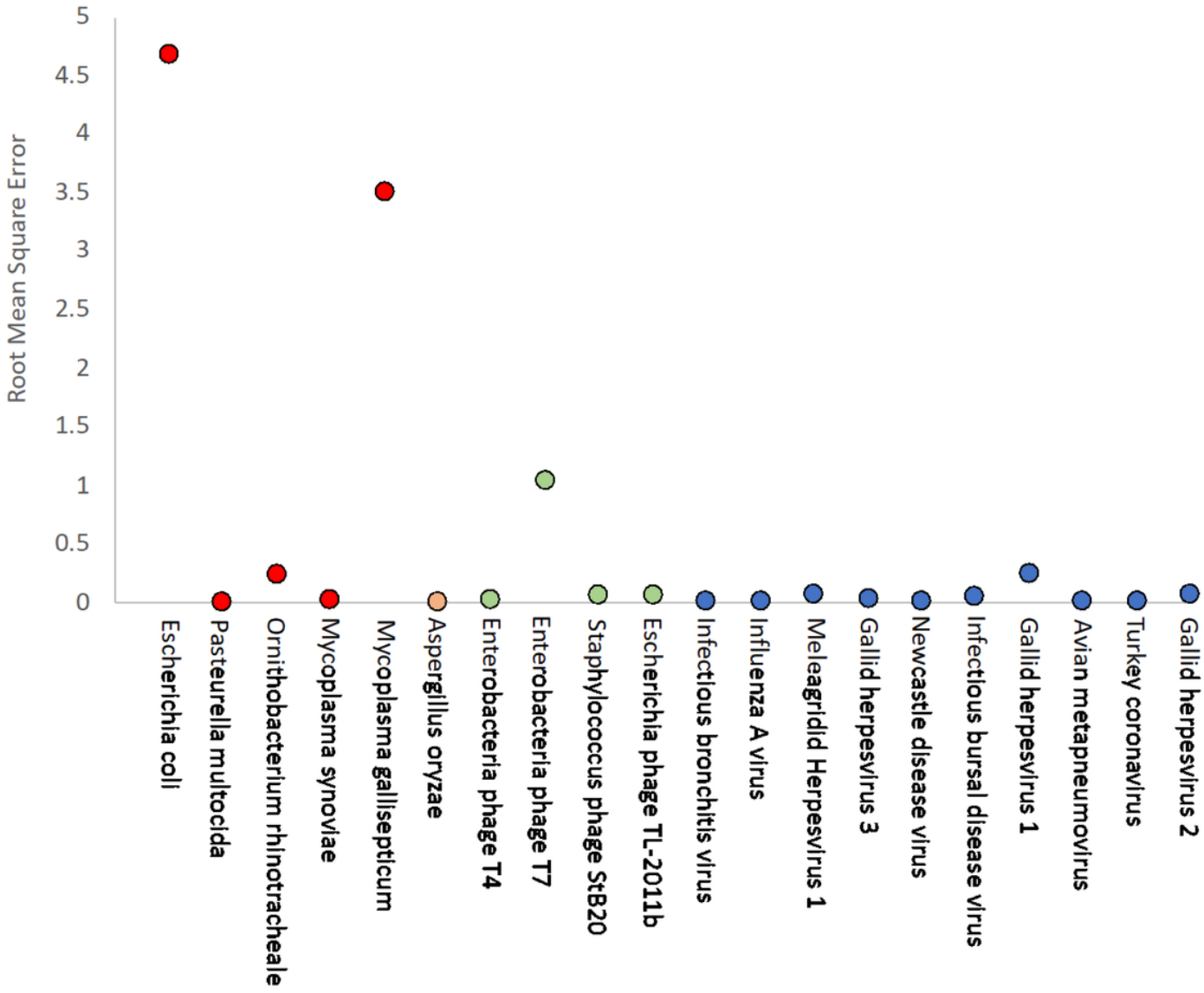

Figure 4

Root Mean Square Error between known abundances and abundances determined by BiomeSeq

Family	Genus	Species	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
Herpesviridae	Iltovirus	Gallid alpha herpesvirus 1	0.171							
	Mardivirus	Gallid alphaherpesvirus 2&3				0.257	0.145	0.070	0.006	
		Meleagrid alphaherpesvirus 1			0.037		0.004			
Anelloviridae	Gyrovirus	Avian gyrovirus					88.664	12.054	15.469	36.773
Adenoviridae	Aviadenovirus	Fowl aviadenovirus							53.299	6.698
Birnaviridae	Avibirnavirus	Infectious bursal disease virus					0.008		2.333	0.105
Coronaviridae	Gammacoronavirus	Avian infectious bronchitis virus	0.382	54.762	58.947	16.278	1.884	23.602	21.786	19.319
Retroviridae	Alpharetrovirus	Avian carcinoma virus						0.077		0.042
	Unclassified	Avian Endogenous Retrovirus	99.447	44.493	41.017	83.296	9.290	64.196	7.108	37.063
Astroviridae	Avastrovirus	Chicken astrovirus		0.744						
Picornaviridae	Sicinivirus	Chicken sicinivirus JSY				0.169	0.005			

Figure 5

Heatmap of percent normalized relative abundance of viruses detected in a commercial poultry flock from hatching to processing. Color corresponds to the range of relative abundance of each family from 0 to 100%. Green: 0-1%; yellow: 1-25%; orange: 25-75%; and red: 75-100%. The sum of each column, or week, is 100%.

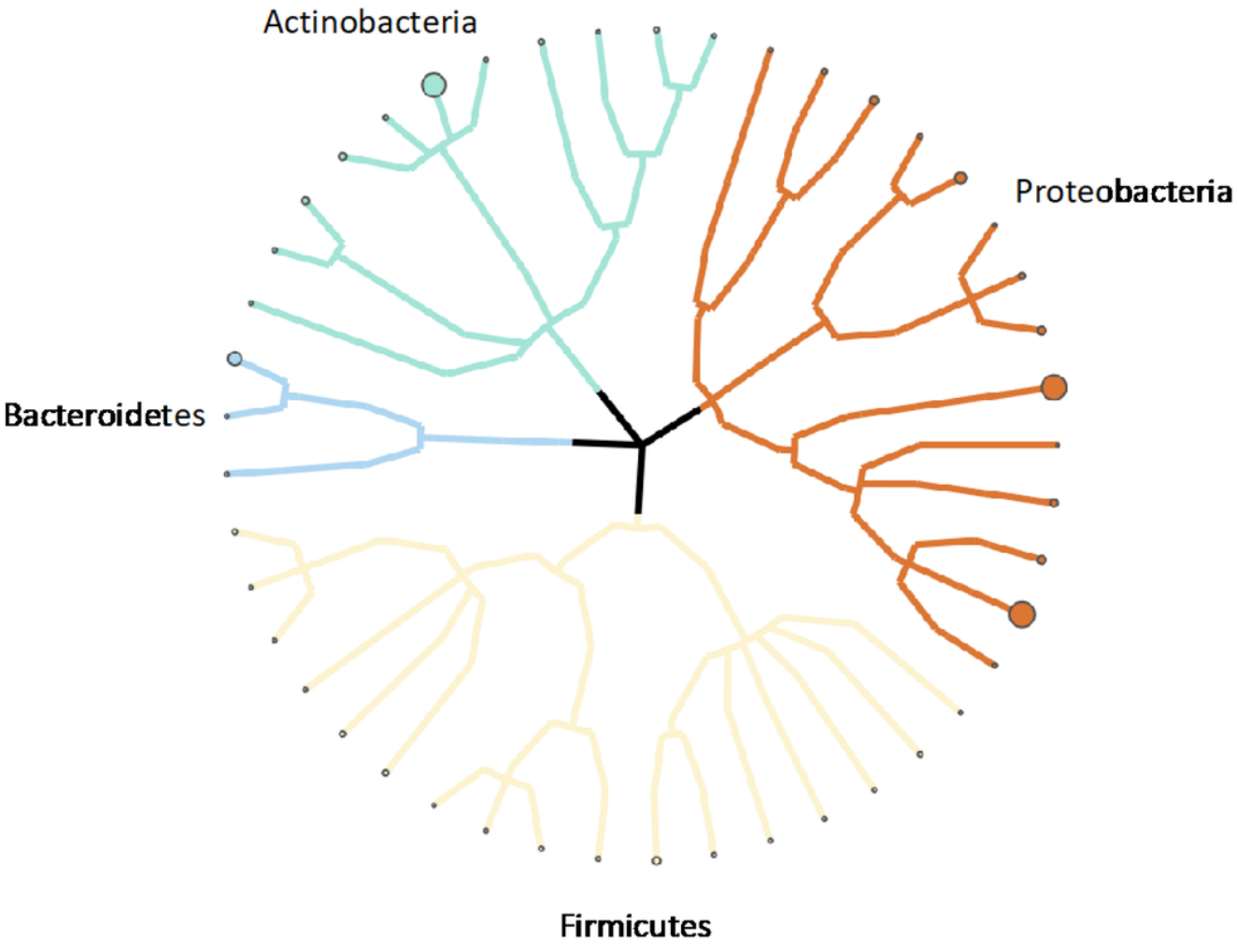

Figure 6

Phylogenetic tree of bacterial species detected in a commercial poultry flock. Branches extend from phylum to species. Nodes indicate detected species and diameter indicates average abundance.

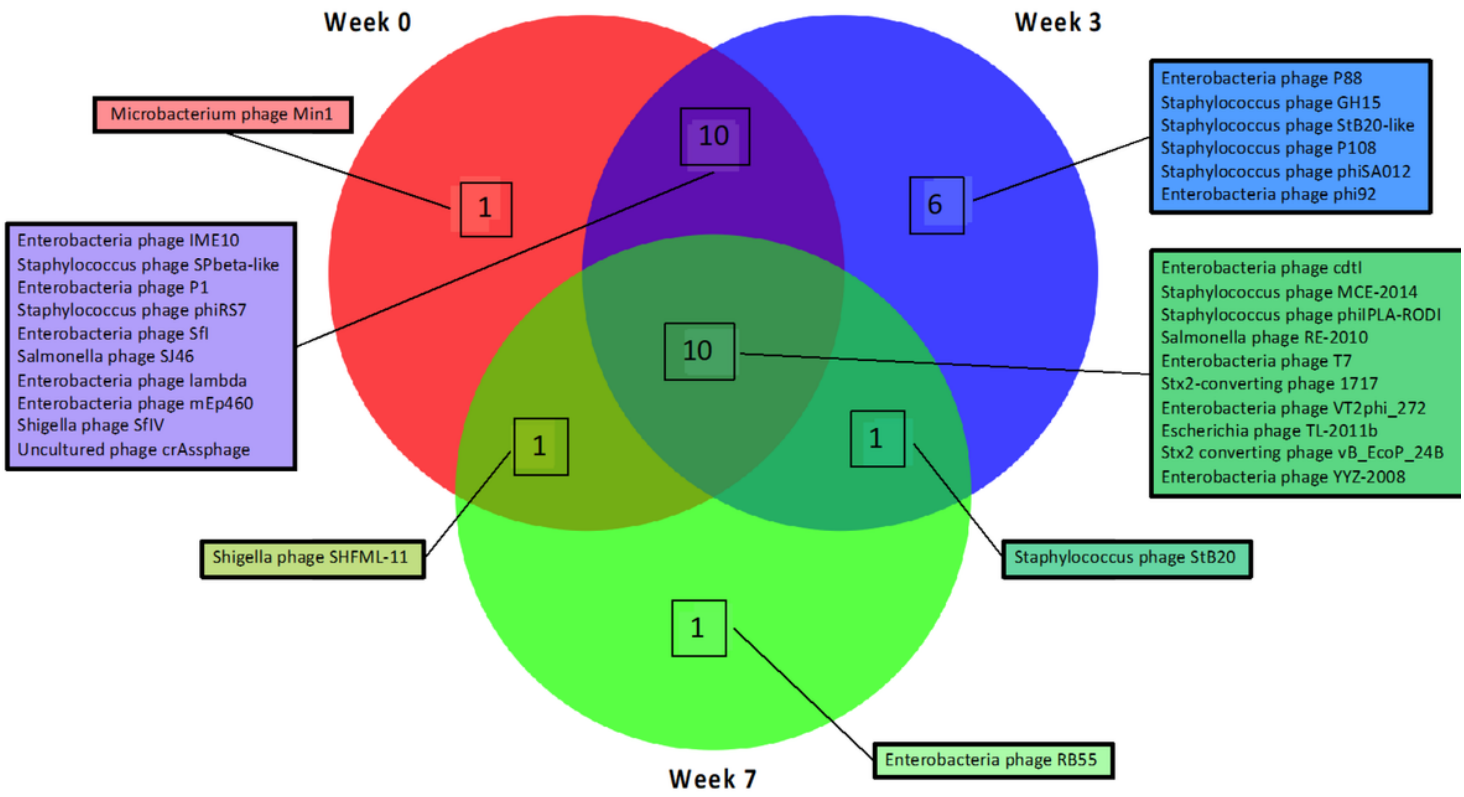

Figure 7

Venn Diagram of the detected bacteriophage species in a commercial poultry flock at Week 0, Week 1 and Week 7.

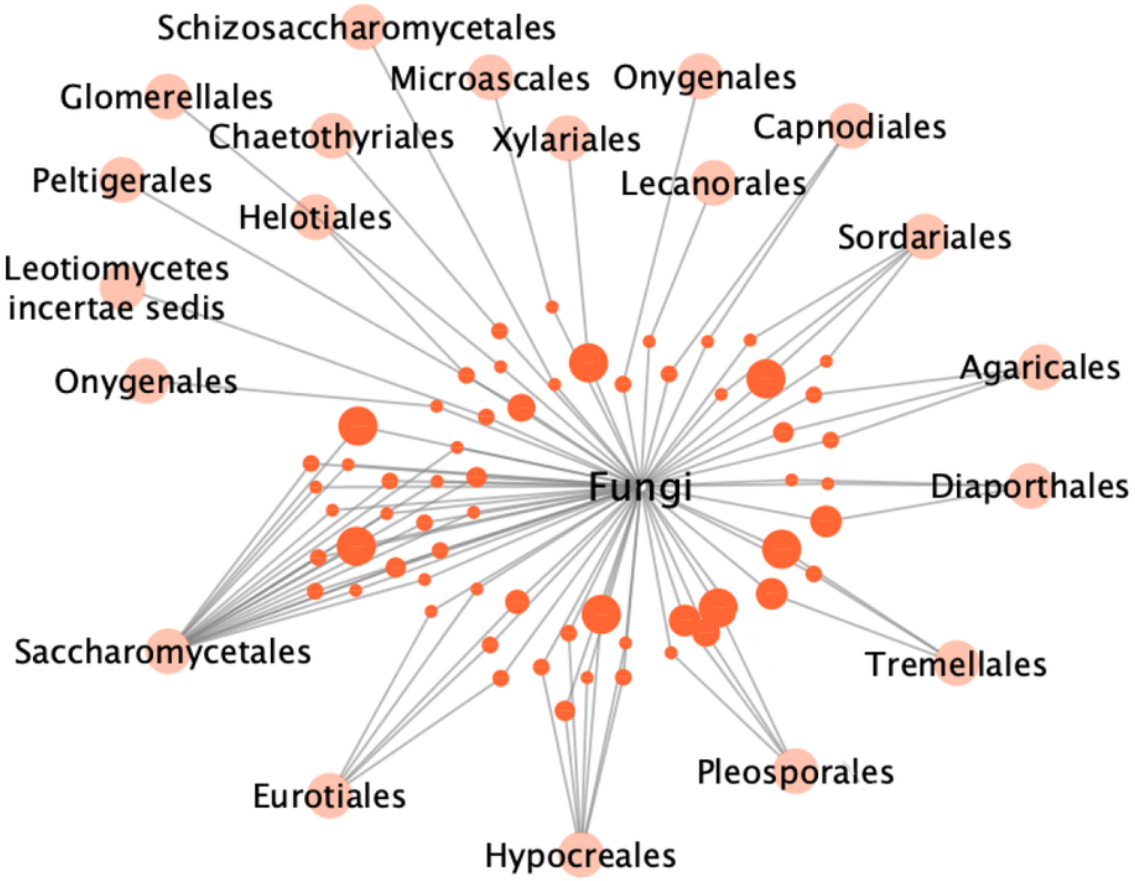

Figure 8

Fungal network of species detected in a commercial poultry flock. Outer nodes represent order level, while inner nodes represent species. The diameter of the inner nodes correlate to species frequency, or the number of weeks the species was detected.

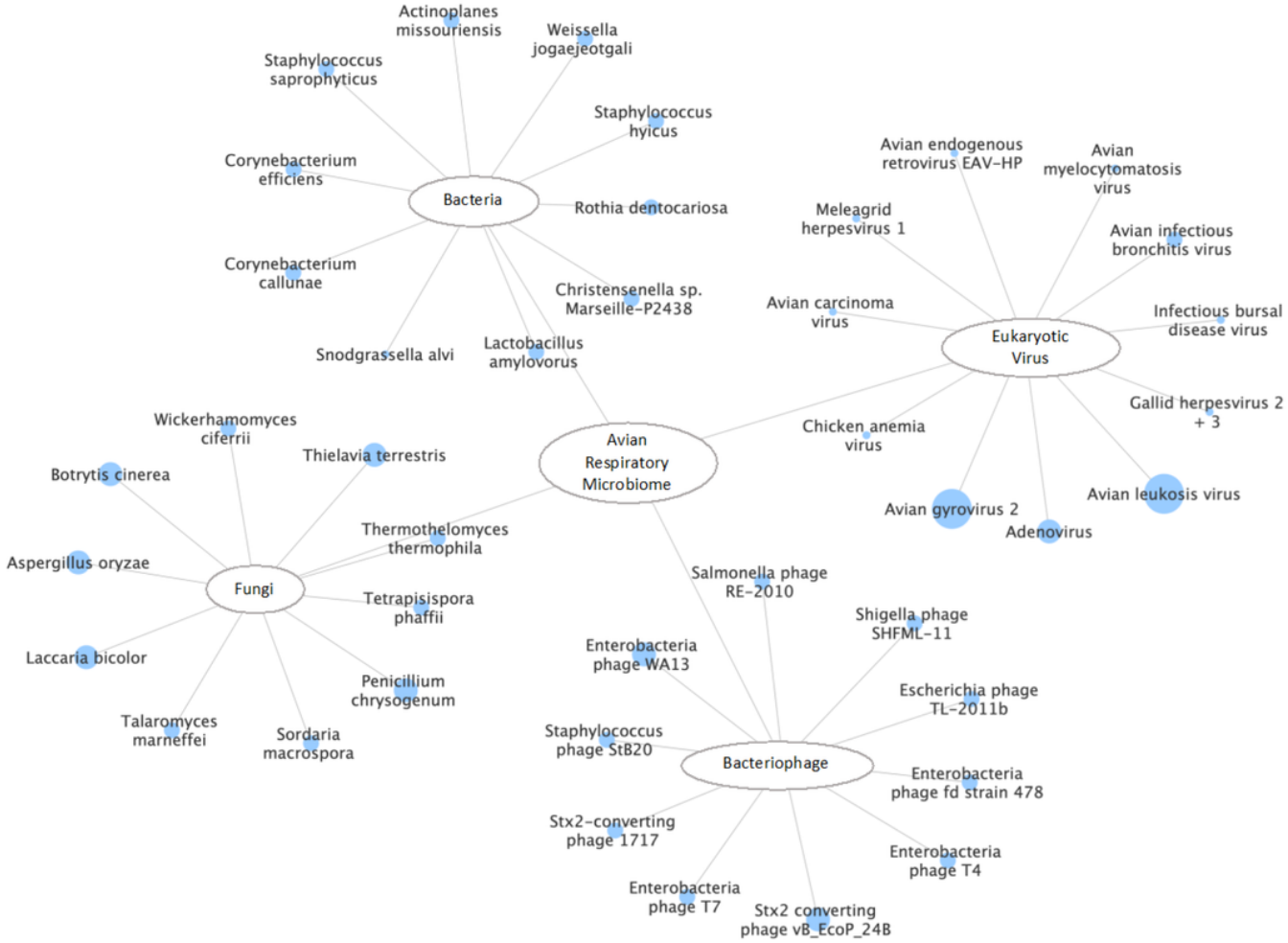

Figure 9

Microbial network of the top 10 most abundant eukaryotic viruses, fungi, bacteria and bacteriophage in a commercial poultry flock at time of processing. Node diameter indicates the percent relative abundance.

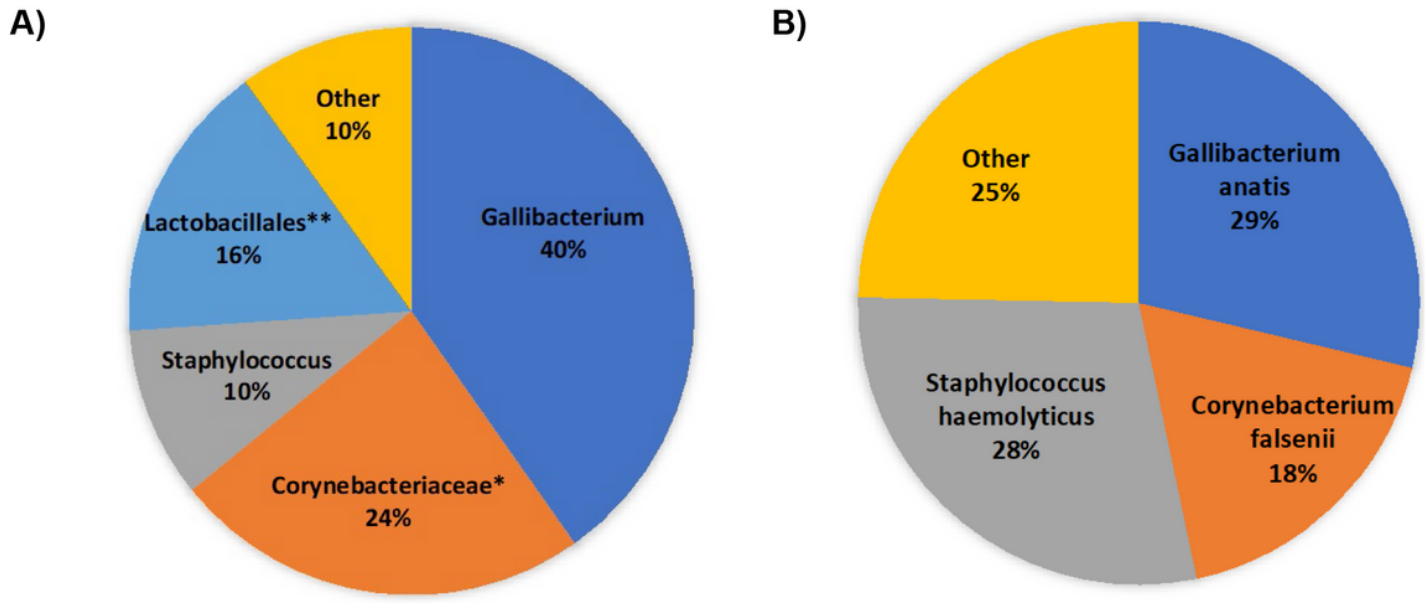

Figure 10

Bacteria detected in a healthy poultry broiler flock using A) 16S rRNA and B) BiomeSeq

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MulhollandSupportingInformation.docx