Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood

Marko Melnick[*], Patrick Gonzales[*], Thomas J. LaRocca[†], Robin D. Dowell[‡],
Yuping Song[**], Joanne Wuu[‡‡], Michael Benatar[‡‡], Björn Oskarsson[§§], Leonard Petrucelli[**,††],
Christopher D. Link[*,§] Mercedes Prudencio[**,††]

[*] Integrative Physiology, University of Colorado, Boulder, Colorado, 80303, USA

[†] Department of Health and Exercise Science, Center for Healthy Aging, Colorado State University, Fort Collins, Colorado, 80523, USA

[‡] BioFrontiers Institute and Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, 80303, USA

[§] Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado, 80303, USA

[**] Department of Neuroscience, Mayo Clinic, 4500 San Pablo Road, Jacksonville, Florida, 32224, USA

[††] Neuroscience Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, Florida, 32224, USA

[‡‡] Department of Neurology, University of Miami, Miami, Florida, 33136, USA

[§§] Department of Neurology, Mayo Clinic, 4500 San Pablo Road, Jacksonville FL, 32224, USA

Data available at NCBI Sequence Read Archive under the accession number (PRJN).
Supplemental material available at figshare: https:// doi.org/(INSERT).
Code available at https://github.com/Senorelegans/MysteryMiner.

Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood

Keywords: ALS, Transcriptomics, RNA-seq, Microbiome, Virome

Corresponding author
Phone: 303-735-5112
Department of Integrative Physiology
354 UCB
Boulder Colorado, 80303
Email: Marko.Melnick@colorado.edu

Mercedes Prudencio
4500 San Pablo Road,
Griffin Building Rm 221
Jacksonville, FL 32224
Phone: 904-953-6638; Fax: 904-953-7370
Email: Prudencio.Mercedes@Mayo.edu

ABSTRACT

Numerous reports have suggested that infectious agents could play a role in neurodegenerative diseases, but specific etiological agents have not been convincingly demonstrated. To search for candidate agents in an unbiased fashion, we have developed a bioinformatic pipeline that identifies microbial sequences in mammalian RNA-seq data, including sequences with no significant nucleotide similarity hits in GenBank. Effectiveness of the pipeline was tested using publicly available RNA-seq data. We then applied this pipeline to a novel RNA-seq dataset generated from a cohort of 120 samples from amyotrophic lateral sclerosis (ALS) patients and controls, and identified sequences corresponding to known bacteria and viruses, as well as novel virus-like sequences. The presence of these novel virus-like sequences, which were identified in subsets of both patients and controls, were confirmed by quantitative RT-PCR. We believe this pipeline will be a useful tool for the identification of potential etiological agents in the many RNA-seq data sets currently being generated.

INTRODUCTION

Background of organisms in neurodegeneration

Infection has been proposed to play a role in multiple neurodegenerative diseases[1], including amyotrophic lateral sclerosis (ALS)[2]. ALS is the most common motor neuron disease in adults, with the majority of individuals dying within 3-5 years of symptom onset. The disease is defined by the degeneration and death of motor neurons in the brain and spinal cord, resulting in progressive weakness and eventually death, typically from respiratory muscle weakness[3]. Around 5-10% of ALS cases are inherited, termed familial ALS (fALS), with the remaining cases considered sporadic ALS (sALS). After decades of study, the etiology of sALS remains a mystery, although suspected risk factors for ALS include exposure to heavy metals, pesticides, chemical solvents, cigarette smoke, and unidentified factors related to US military service[4–7].

Along with these environmental risk factors, there has been a long history, with variable success, in the search for pathogens that might contribute to ALS[8–12] and other neurodegenerative diseases such as Alzheimer's disease (AD)[13–15], Parkinson's disease (PD)[16–18], and multiple sclerosis (MS)[19].

Diverse pathogens have been reported in the blood, cerebrospinal fluid (CSF) and central nervous system (CNS) from ALS patients. For example, bacteria that have been detected include *Cutibacterium acnes*, *Corynebacterium sp*, *Fusobacterium nucleatum*, *Lawsonella clevelandesis*, and *Streptococcus thermophilus* in CSF[20], and mycoplasma in blood[21]. Fungi, including *Candida famata*, *Candida albicans*, *Candida parapsilosis*, *Candida glabrata*, and *Penicillium notatum*, have been detected in CSF, while *Malassezia globosa*, *Cryptococcus neoformans*[11], and *Candida albicans* have been found in various regions of the CNS[11,22,23]. The search for viruses that contribute to ALS pathology is much more extensive and includes studies on herpes virus[9,24], enterovirus[9,25–28], human immunodeficiency virus (HIV)[29,30], and human endogenous retrovirus (HERV-K)[31–33]. Importantly, multiple studies using immunohistochemistry have shown an increased load of various pathogens in ALS samples compared to controls in multiple tissues suggesting these pathogens are present and cannot be simply attributed to contamination[9,11,20,22,23]. Ultimately, the presence of ALS dysbiosis is unresolved and remains an active area of investigation with evidence for[34–38] and against[39] it.

The biological role that these alternative microbiotas play in ALS is also unclear. ALS patients may have a compromised blood brain barrier (BBB) or blood spinal cord barrier (BSCB) function[40,41]. It has been reported that ALS patients also have elevated Gram negative endotoxin/lipopolysaccharide (LPS) in the blood[42]. Patients with ALS also display activation of the innate immune system along with changes in blood[43,44], spinal cord and motor neurons[45], but if and how bacteria might influence activation is an active area of research. A “dual hit” hypothesis by Correia et al. suggests inflammation via LPS may contribute to mis-localization and aggregation of ALS-implicated protein TAR DNA-binding protein 43 (TDP-43)[46].

Numerous studies have looked for biomarkers of ALS[47] using metabolomics[48,49], neuroinflammation[50,51] , DNA methylation[52,53], gene expression[54], microRNA expression[55,56] and our previous study which analyzed protein levels of poly(GP) in *C9ORF72*-associated ALS (c9ALS)[57]. The search for pathogens using sequencing data from blood samples in ALS patients has been conducted before[58–61], but previous efforts have not looked for novel pathogens.

Next-generation sequencing (NGS) technologies have shown broad detection of pathogens in a target-independent unbiased fashion[62–65]. However, designing a microbial detection experiment is non-trivial considering the variety of methods[66] and algorithms[67] that can be applied. Our primary goal when designing a new pipeline was to be conservative and unbiased with regards to discovery of novel pathogens. Furthermore, we wanted our pipeline to allow for the quantification of both novel and known pathogens. While other pipelines have used reads that do not map to the host genome (unmapped reads) for microbial identification and quantification, these pipelines cannot be used for discovery as they rely on existing databases of microbial genomes[68–71]. Thus, we opted for de-novo assembly of unmapped reads into contigs, followed by alignment of unmapped reads back to these contigs for quantification. A similar pipeline known as IMSA[72] uses this strategy, but disregards contigs that might be identified by translated amino acid sequence similarity using BLASTX (a set we call the “dark biome”) as well as subsequent contigs with no BLASTN or BLASTX hit (a set we call the “double dark biome”).

We validated our pipeline by using datasets with known bacterial or viral infections. We also examined the differences in microbial identification between polyA and total RNA recovery in multiple tissues, and investigated the effects of globin depletion of blood samples. We then used our pipeline on a novel blood dataset (termed “Our Study”) as well as on five other published ALS datasets from blood or spinal cord samples, analyzed each dataset individually, and analyzed across datasets for changes in microbiota. While we did not identify any microbes enriched in the blood of ALS patients, we did identify and validate a novel virus-like sequence, demonstrating the potential of the bioinformatic pipeline we have established.

MATERIALS AND METHODS

Blood Collection and RNA Extraction

A total of 120 RNA whole blood samples constitute Our Study, which included 30 healthy controls (from general population that do not have blood relatives suffering from ALS, CTL), 30 pre-symptomatic *C9ORF72* mutant carriers (C9A), 30 symptomatic *C9ORF72* ALS cases (C9S), and 30 symptomatic *C9ORF72*-negative ALS cases (SYM). PAXgene blood RNA tubes were collected at Mayo Clinic Jacksonville and at University of Miami. All 120 RNA samples selected for RNA-seq were received and processed at Mayo Clinic Jacksonville using PAXgene blood RNA kit following manufacturer's recommendations (Qiagen). Blood RNA was of high quality, assessed in an Agilent Bioanalyzer (Agilent), with RNA integrity values ranging from 7.4 to 9.8, with a median value of 8.7. RNA samples were then sent to The Jackson Laboratory for globin depletion, library preparation and sequencing of total blood RNA.

Globin Depletion

Due to the abundance of large haemoglobin RNA transcripts present in the blood, a globin depletion step, using the Ambion GLOBINclear kit (AM1980), was performed before sequencing of the blood RNA samples in order maximize coverage on non-globin genes. In brief, one microgram of total RNA was used as starting material, and specific biotinylated oligos were used to capture globin mRNA transcripts. The capture oligos were hybridized with total RNA samples at 50°C for 30 min. Streptavidin magnetic beads were then used to bind to the biotinylated capture oligos hybridized to globin mRNA by incubating at 50°C for 30 min. The magnetic streptavidin beads-biotin complex were then captured to the side of the tubes by a magnet, and the resulting supernatant is free of globin mRNA. The globin depleted RNA was further purified by RNA binding beads and finally eluted in elution buffer. The resulting RNA free of >95% globin mRNA transcripts was then processed for next generation sequencing. Of note, to assess the efficiency of the globin RNA depletion, 10% of the samples processed were selected randomly and amplified using a Target-Amp Nano labeling kit (Epicentre). Samples were normalized to 100 ng input and reverse transcribed. First strand cDNA was generated by incubating at 50°C for 30 min with first strand premix and Superscript III. This was followed by second strand cDNA synthesis through DNA polymerase by incubating at 65°C for 10 min and at 80°C for 3 min. In-vitro transcription was then performed at 42°C for 4 hours followed by purification using RNeasy mini kit (Qiagen).

Due to the large number of samples, the globin depletion step was performed in two batches. We provided guidelines on how samples would be divided among the batches and also for how samples would be grouped in the sequencing runs in order to minimize technical variability. The Jackson Laboratory personnel were blinded to the identity of the samples.

RNA-seq of total blood RNA (globin and ribosomal RNA depleted) was performed in an Illumina HiSeq4000 with >70 million read pairs per sample. Raw reads were then sent back to us for bioinformatics analyses.

Quantitative RT-PCR for blood RNA samples

A total of 500 ng of total blood RNA was used for reverse transcription polymerase chain reaction (RT-PCR), using the High Capacity cDNA Transcription Kit with random primers (Applied Biosystems). Quantitative real-time PCR (qRT-PCR) was performed using SYBR GreenER qPCR SuperMix (Invitrogen). Samples were run in triplicate, and qRT-PCRs were run on a QuantStudio 7 Flex Real-Time system (Applied Biosystems).

List of primers and their sequences:
RDRP forward 5'-GCTGTCAAATCGGTTTCCAAC-3';
RDRP reverse 5'-CTGCCTTCGTCATCTTGGAG-3';
GAPDH forward 5'-GTTCGACAGTCAGCCGCATC-3';

GAPDH reverse 5’-GGAATTTGCCATGGGTGGA-3’.

Transcriptomics

See pipeline description in results for an overview of the pipeline; see bioinformatics supplement File S1 for a more detailed description of the analysis pipeline, versions, and statistical quantification. All data in this study was processed identically using the pipeline.

Statistical Analysis

To assess statistical differences between conditions, a two tailed Student’s *t*-test is calculated using normalized coverage for contigs or binned normalized coverage for species/genus, etc. The number of contigs or genus/species is used to obtain an adjusted p-value using scipy in Python. Cutoff for statistical significance is less than an adjusted p-value of 0.05 unless otherwise stated.

Data availability

Raw sequencing data for Our Study dataset is available in the NCBI Sequence Read Archive under the accession number (PRJN). All other datasets are publicly available and all of the code used in this manuscript is available at https://github.com/Senorelegans/MysteryMiner. Supplemental material available at figshare: https:// doi.org/(INSERT).

RESULTS

Pipeline description

Mystery Miner is written as a Nextflow pipeline. Below is a short overview of the Mystery Miner pipeline (Fig1). A more in-depth explanation, list of software and versions used, and typical parameters of each step is described in the bioinformatics supplement, and all of the code used in this manuscript can be found at https://github.com/Senorelegans/MysteryMiner.

Raw reads were first checked for quality using FastQC then trimmed to remove both adaptor contamination and low quality basecalls using Trimmomatic. Trimmed reads were then mapped to the host genome using multiple alignment algorithms in series (STAR, Bowtie2) and unmapped reads were retained for contig assembly. Filtering out host reads made downstream assembly faster and required less memory. We assembled contigs from unmapped reads with the SPAdes assembler (with “-rna” setting). This assembler was chosen for its recent use in studies of microbial diversity[73] and proven robustness to biological and technical variation[74]. The species each contig belongs to was identified with BLASTN using default settings, and the top hit for each contig was retained (a set we call “regular biome”). Contigs with no BLASTN hits were then filtered to remove highly repetitive regions (DUST) and retained if they had a greater than 60% pairwise alignment (LAST) between contigs assembled from a single sample, group/condition, or all samples.

We then identified contigs that lacked detectable nucleotide similarity to any GenBank entry but showed similarity at the amino acid level using BLASTX (“dark biome”). Furthermore, contigs with no BLASTN or BLASTX hits were labelled as “double dark biome” (also filtered by

DUST and LAST). Every contig in the "regular biome" and "dark biome" were then queried against the Joint Genome Institute Server for additional taxonomic information. As Mystery Miner is an agnostic tool, it cannot distinguish between true tissue or cell-associated microbes and experimentally introduced contamination.

For quantification, we mapped the non-host reads using Bowtie2 to the contigs obtained from SPAdes. Next, we mapped reads to contigs using samtools mpileup (default mapq score) to calculate the amount of reads over each base pair in a contig. We then calculated coverage on the contigs by summing all of the counts for each base pair in a contig and dividing by the length of the contig. We then calculated normalized coverage by library size using the number of mapped reads to the host genome. This gave us normalized coverage (NC) for a contig or binned normalized coverage (BNC) for multiple contigs within a species/genus, etc. To assess statistical differences between conditions, a Student's *t*-test was calculated through NC or BNC, using the number of contigs or genus/species to obtain an adjusted p-value using scipy in Python.

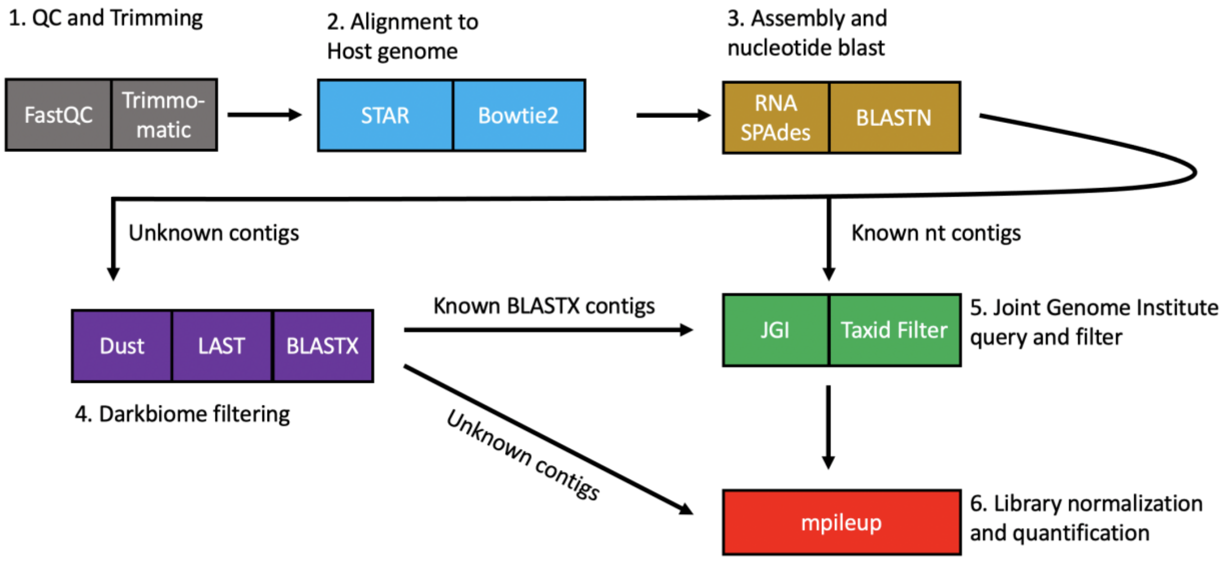

Figure 1. Diagram of Mystery Miner Pipeline
Reads were first checked with FastQC and trimmed using Trimmomatic (1. grey). Reads were then aligned to the host genome using various aligners (2. blue). Non-host (unmapped) reads were assembled into contigs with RNA SPAdes and regular biome contigs were identified with BLASTN (3. yellow). Unidentified contigs were filtered for repetitive sequences with Dust, filter by single, group or all with LAST, and dark biome contigs were identified with BLASTX. Double dark biome unidentified BLASTX contigs were sent directly to quantification (4. purple). Dark biome and regular biome contigs were assigned complete taxonomy using the JGI server and filtered one last time to remove mammalian/host genome contigs (5. Green). Non-host reads were then mapped to all contigs and normalized coverage was calculated for subsequent statistical analysis.

Validating Mystery Miner on datasets with known bacterial or viral infection

To confirm that Mystery Miner is able to recover and quantify known bacterial infections from sequencing data, we utilized an *in vitro* model of *Chlamydia trachomatis* infection (Humphrys 2016)[75]. In this study, epithelial cell monolayers were infected with *Chlamydia trachomatis*; and polyA RNA (6 samples) and total RNA (6 samples) were sequenced 1 hour and 24 hours post infection (hpi). Using the Mystery Miner pipeline, out of 5.32 X 10^6 reads from all of the samples, 6.04 X 10^5 reads remained unmapped (~11.34%) after trimming and mapping to the host genome (File S2). From these non-host reads, 3,257 contigs were assembled and 1,199 of these contigs were identified as regular biome (File S3). An additional 27 contigs had no BLASTN hit. Of these, we identified 2 dark biome (BLASTX identified) and no double dark biome (no BLASTX or BLASTN hit) contigs (File S4 and File S5).

Using Mystery Miner we successfully identified, and found significantly elevated levels, of *Chlamydia trachomatis* (BNC by species) in 24 hours post infection (hpi) samples compared to 1 hpi samples in both polyA (Padj = 0.004) and total RNA (Padj = 0.0005). In addition to *Chlamydia trachomatis*, we identified 6 additional bacterial species and one viral species (Alphapapillomavirus 7) in the samples (Figure 2A), including significantly elevated levels of *Mycoplasma hyorhinis* contigs in total RNA samples. No significant differences were observed in the dark or double dark contigs (File S6).

To confirm that the pipeline can detect known viral infections, we ran Mystery Miner on a total RNA dataset from an *in vitro* model of severe acute respiratory syndrome coronavirus (SARS-CoV) 1 or 2 infection (Emanuel2020[76]). In this study human epithelial Calu3 cells were infected with SARS-CoV-1 or SARS-CoV-2 (4, 12, or 24 hours), mock (4 or 24 hours), or untreated (4 hours).

Out of the 2.81 X 10^8 reads obtained from all of the samples, 8.23 X 10^7 reads remained unmapped (~29%) after trimming and mapping to the host genome (File S2). From these non-host reads, 42,816 contigs were assembled, of which 1,346 regular biome, 27 dark biome, and 7 double dark biome contigs passed the filtering steps (File S2, File S3, File S4, File S5)

Mystery Miner successfully identified both SARS-CoV-2 and SARS-CoV-1 isolates and found significantly elevated levels of each virus compared to controls (Figure 2B). Hereafter we refer to SARS-CoV-1 or SARS-CoV-2 infected cells as COV1 or COV2 to avoid confusion with recovered names of contigs. Consistent with the viruses being similar, we identified both SARS-CoV-2 and SARS-CoV-1 in both the COV1-24hr and COV2-24hr samples when compared to mock-24hr. However, when we compared COV2-24hr to COV1-24hr, we were able to distinguish SARS-CoV-1 isolates from SARS-CoV-2 in the appropriate samples (i.e., SARS-CoV-2 was significantly elevated in COV2). Similar results were seen in the 12 hour samples but the 4 hour samples did not have sufficient viral reads to detect either SARS-CoV virus (File S7). To simulate a novel virus, we ran Mystery Miner on versions of the BLASTN and BLASTX databases obtained before SARS-CoV-2 was discovered and were able to properly identify SARS-CoV-2 as a bat related coronavirus[77] (Figure S1) (File S7).

Collectively, these data show that Mystery Miner is able to identify potential bacterial and viral infections, properly identify infected samples using quantification, and detect significant differences between infected samples and controls for bacteria, viruses, and isolates of a virus.

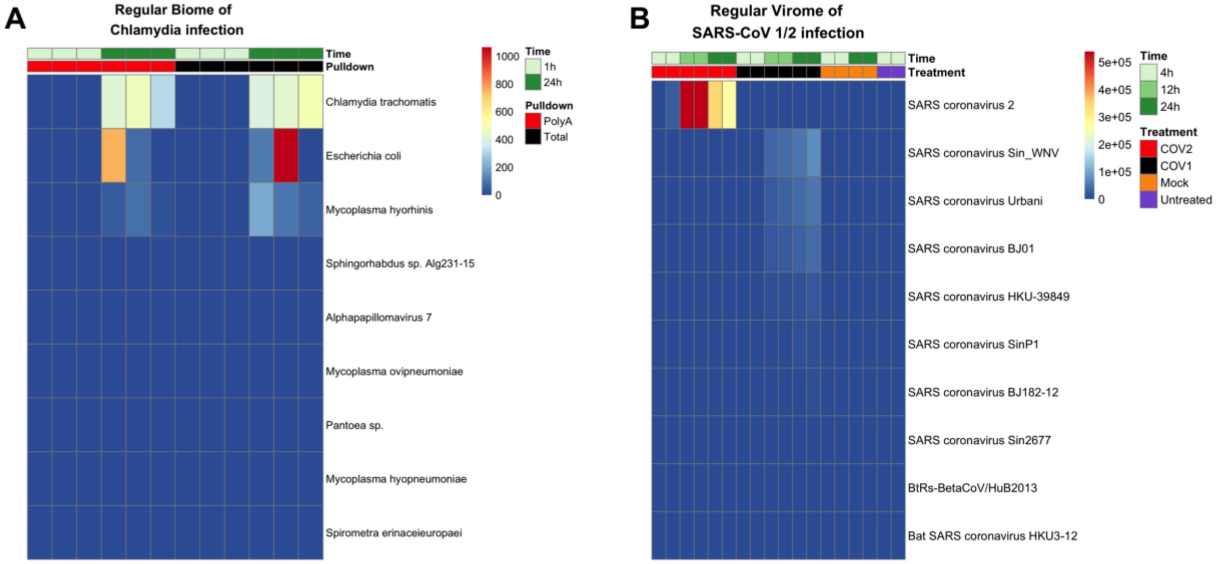

Figure 2. Heatmap of binned normalized coverage for bacterial or viral infected datasets. A. Regular biome contigs binned by species from Humphrys et al., 2016. Time refers to 1or 24 hours post infection (hpi) of epithelial cell monolayers with *Chlamydia trachomatis* (green). Pulldown refers to library enrichment for polyA RNA (red) or total RNA (black). **B**. Regular virome of contigs binned by name from Emanuel et al., 2020 for SARS-CoV-2 infected cells (COV2) (red), or SARS-CoV-1infected cells (COV1) (black), mock virus (orange), or untreated sample (purple). Time refers 4,12, or 24 hpi of Calu3 cells with indicated virus (green). Top 10 hits per experiment shown for brevity.

Effects of library pulldown or globin depletion in RNA-seq datasets

In order to compare effects of library enrichment or depletion, we compared recovered pathogens in a dataset that has polyA enrichment or rRNA depleted total RNA from blood or colonic tissue (VonSchack2018)[78]. When we compared polyA RNA vs total RNA and looked at BNC by superkingdom of bacteria we found significantly more reads map to bacteria for total RNA than polyA RNA (Padj = 0.0349), in blood but not in colon (Padj=0.11709) (Figures S2 and File S8). We found similar amounts of significant BNC by species for polyA RNA vs total RNA in blood (29) and in colon (26). We then looked at significant BNC by genus and found double the amount in blood (14) compared to colon (7), with only one significant genus (*Actinomyces*) found in both comparisons. We did not find any significant differences in coverage when we looked at the species, genus or superkingdom level for viruses (File S8). We conclude that library enrichment with total RNA compared to polyA RNA increases bacterial recovery and diversity in blood.

We next looked at a RNA-seq dataset from whole blood with globin depleted (GD) vs non-globin depleted (NGD) total RNA (Shin2014[79]). With BNC by superkingdom, we found significantly increased levels in globin depleted vs. not-depleted samples for both bacteria (Padj = 0.004) (Figure S3) and viruses (Padj = 0.030) (Figure S4). We also found significant differences in BNC by species (Figure S5) or genus (Figure S6) primarily from *E. coli* with elevated levels in globin-depleted blood RNA. We did not find any significant differences when we looked for viruses at the species or genus level (File S9).

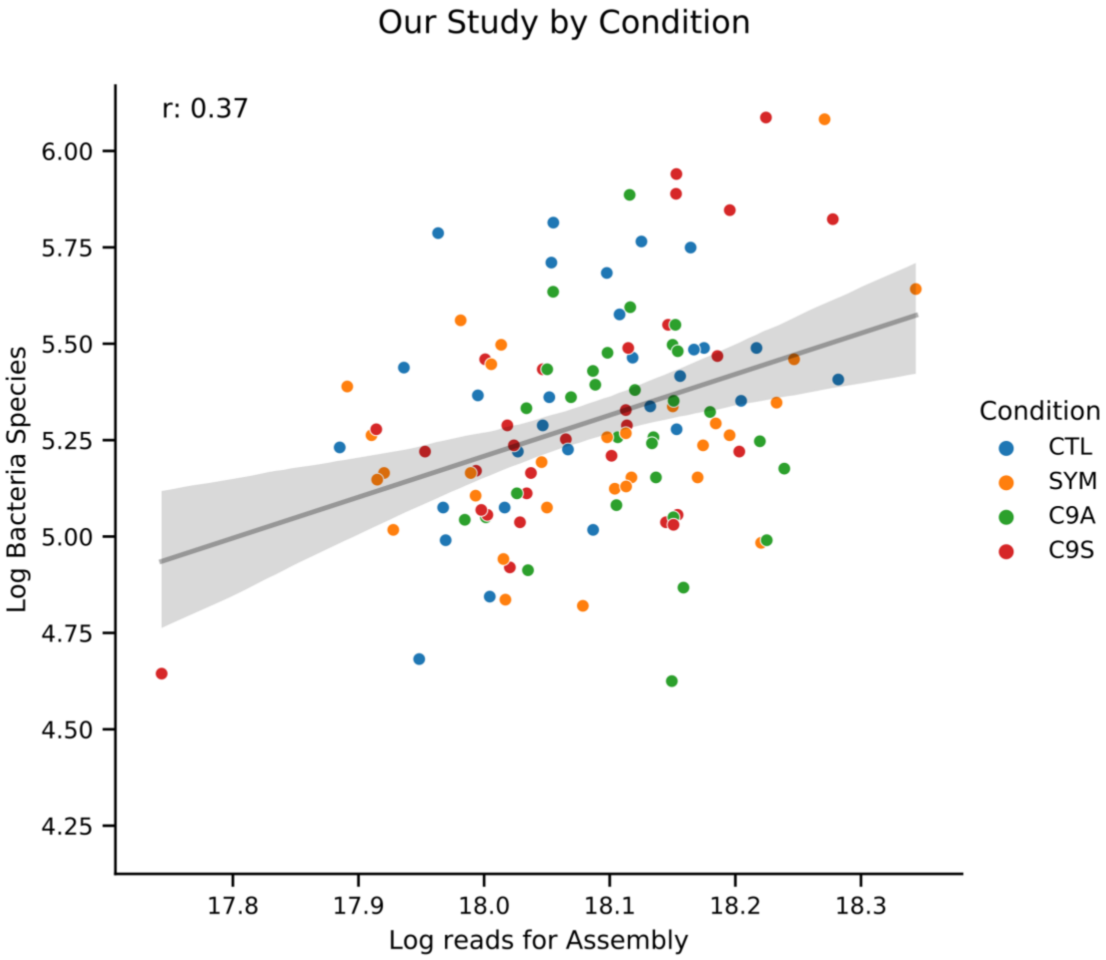

Figure 3. Log number of bacterial species vs Log reads for Assembly in Our Study. Scatterplot where each dot is a sample from a dataset with log number of bacterial contigs assembled on the Y-axis and Log reads used in SPAdes on the X-axis. Samples show a modest correlation (Pearson's r=0.37) between library size and bacterial species recovered. Data fit with a regression (black line) and 95% confidence interval (gray area).

Analysis of Our Study

We used Mystery Miner on our novel RNA-seq dataset of globin depleted and rRNA depleted total blood RNA from 120 individuals. These samples were from four subject groups including healthy control participants (CTL), ALS symptomatic *C9ORF72* negative patients (SYM), *C9ORF72* positive ALS symptomatic patients (C9S) and *C9ORF72* positive asymptomatic individuals (C9A).

The entire dataset contains a combined 8.64 X 10^9 reads. Approximately 2.7% (2.34 X 10^8) of the reads did not map to the human genome. From these non-host reads 2,976,988 contigs were assembled and 17,047 BLASTN contigs (regular biome) were identified. A total of 25,815 contigs had no BLASTN hit and after filtering we identified 2,980 dark biome (BLASTX identified) and 859 double dark biome (no BLASTX or BLASTN hit) contigs (File S2, File S3, File S4, File S5).

In general, we found a modest positive correlation between library size and number of bacterial contigs assembled, species detected (Figure 3), and genera detected for each sample as well as a homogenous mixture of samples with respect to disease status. No specific taxonomy or contig sequence correlated with participant class within the dataset. Yet, by pooling bacterial read counts across all of the samples, we found *alpha proteo-bacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* as the most highly represented taxonomies, consistent with other blood biome studies[80] (Figure S6). Most of the bacterial genera (~65%) assigned to the dark biome contigs were also found in the regular biome, however this was not the case in the viral sets, as only 5% (4/78) of dark viral contigs were observed in the regular biome (File S10). This observation suggested that our pipeline might be identifying novel viral sequences.

Within the dark biome contigs, we noted numerous contigs with a region of protein sequence similarity to RNA-dependent RNA polymerase (RdRP) from multiple RNA viruses, showing highest similarity to the velvet tobacco mottle virus (first row in heatmap of Figure 4, complete metadata shown in Figure S7). Our attention was drawn to the largest (~5 kb) dark biome contig (one of the contigs showing similarity to the velvet tobacco mottle virus) hereafter labeled as “RDRP contig”. To confirm the presence of the RDRP contig in the original samples, we designed primers to the RDRP contig and performed reverse transcriptase polymerase chain reaction (RT-PCR) on seven samples, four of which had high coverage (predicted present) and three with zero coverage (predicted absent). We found typical levels for detection of a virus[81] in the samples with high coverage and detected nothing in samples with zero coverage (Table 1). We conclude that Mystery Miner is biologically validated and can recover unknown pathogens from human subjects.

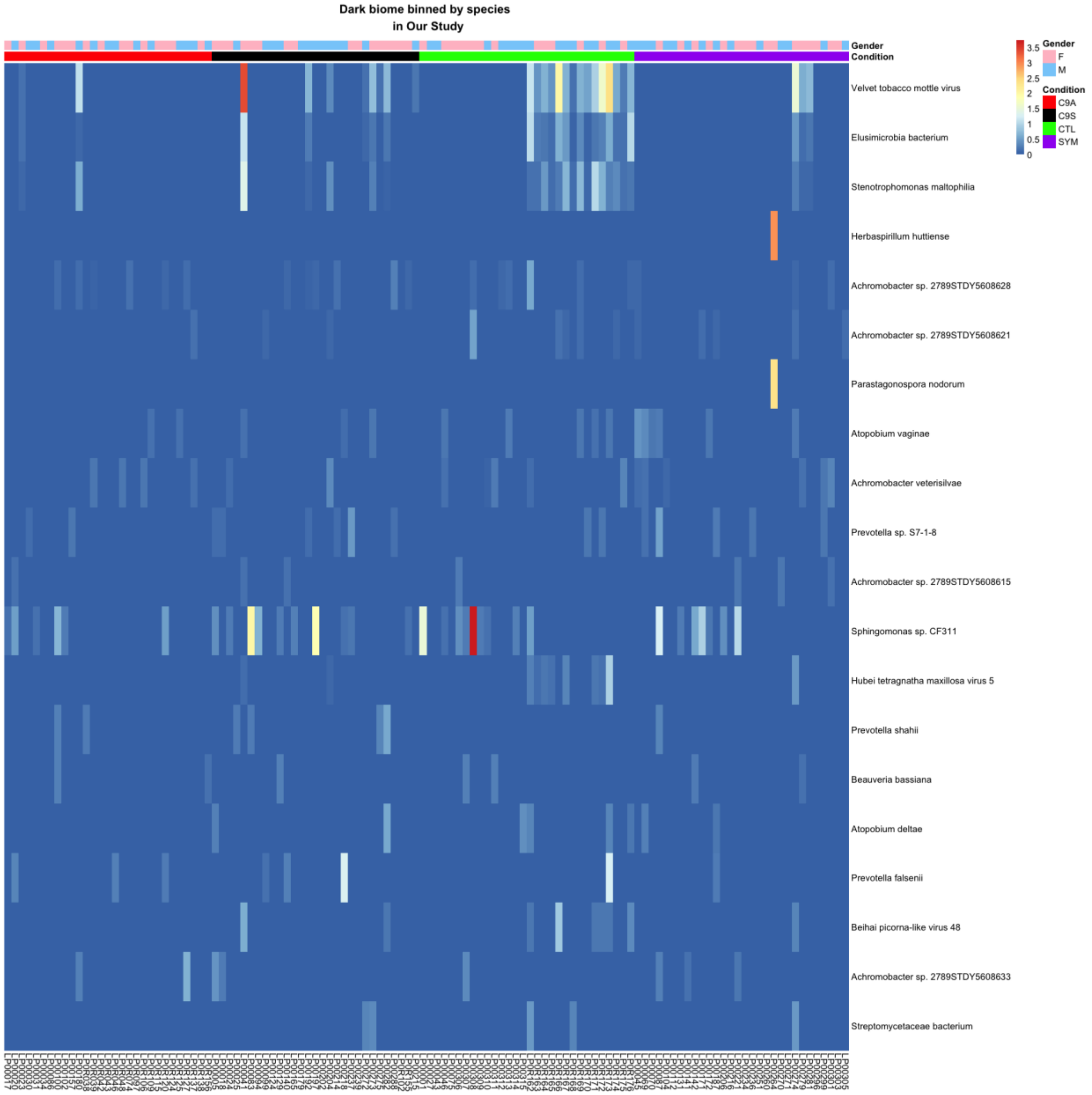

Figure 4. Heatmap of dark biome contigs binned by species in Our Study.
Heatmap of normalized coverage of dark biome contigs binned by species. The highest coverage belongs to contigs that show high similarity to velvet tobacco mottle virus. Zero coverage is blue and goes to red with increasing values. These samples were from four subject groups including healthy controls [(CTL) green], *C9ORF72* negative ALS symptomatic [(SYM) purple], *C9ORF72* positive ALS symptomatic [(C9S) blue] and *C9ORF72* positive asymptomatic [(C9A) red] patients. Sex indicated as light blue (male) and pink (female). Top 20 species sorted by binned normalized coverage was shown for brevity.

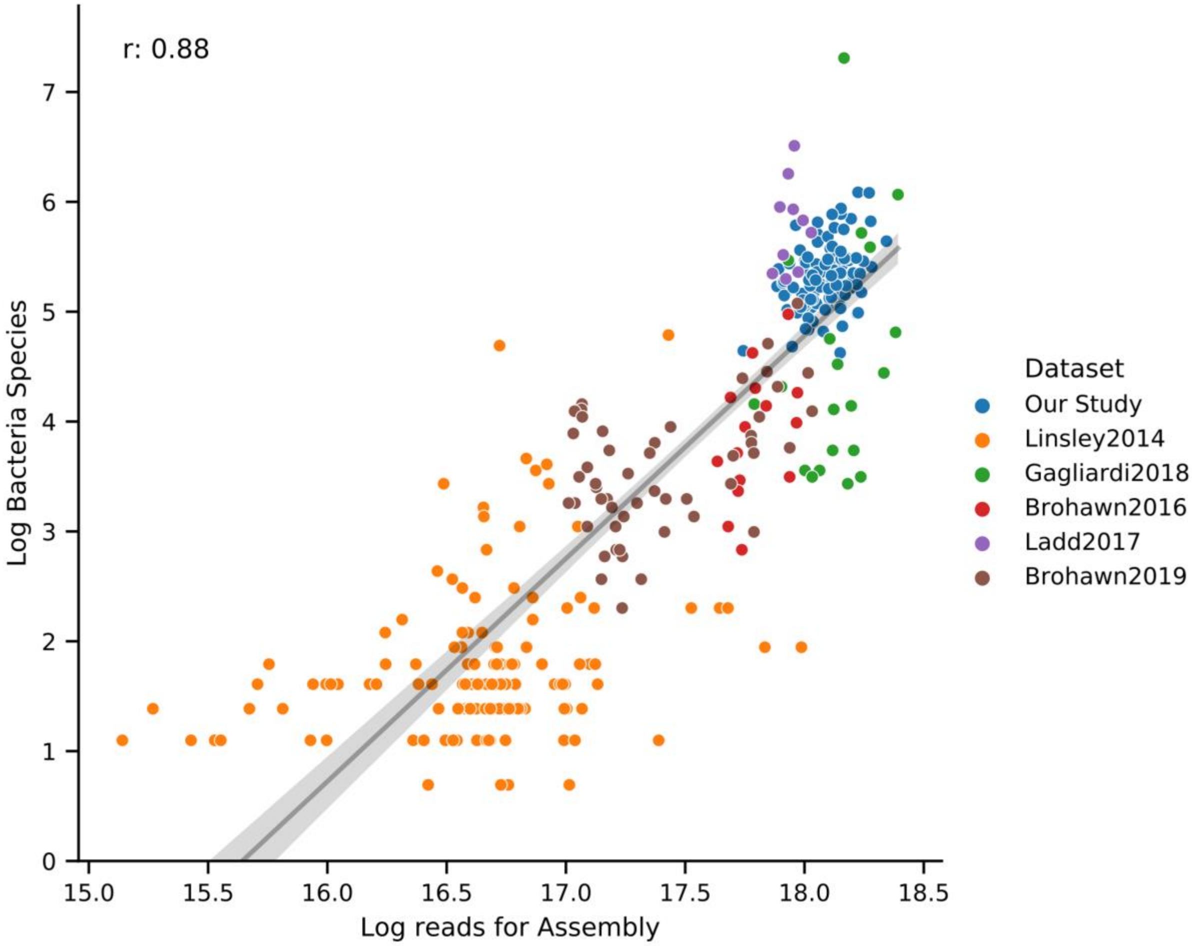

Figure 5. Log number of bacterial species vs Log reads for Assembly for ALS Datasets. Scatterplot where each dot is a sample from a dataset with log number of bacterial contigs assembled on the Y-axis and Log reads used in SPAdes on the X-axis. ALS datasets show a high correlation (Pearson's r=0.88) between library size and bacterial species recovered. Data fit with a regression (black line) and 95% confidence interval (gray area).

Analysis of published ALS datasets

We next sought to explore whether similar results would be obtained from other ALS datasets. To this end, we examined five other publicly available ALS datasets, consisting of two that used total RNA from blood (Linsley2014[82], Gagliardi 2018[58]), and three datasets from spinal cord (Brohawn2016[83], Ladd2017[84], Brohawn 2019[85]). We have provided a summary table of datasets for all studies used in this paper (Table 2). As we observed in Our Study, we first noted that increased library size correlated with an increased number of bacterial contigs assembled, species detected, and genera detected (Figure 5, and Figure S8-10 show all datasets used in this study).

We then looked at the total overlap of genus or species to determine if there are similarities in recovered microbial or viral sequences between datasets. For genus in the regular bacteriome, our dataset had the highest number of unique genus (185), followed by Ladd2017 (117), and Gagliardi2018 (38). The highest number of overlapping bacterial genus was between our dataset and Ladd2017 (133) followed by the intersection between our dataset, Ladd2017 and Gagliardi2018 (61) and there was a modest overlap (24) for 9/10 datasets (Figure 6). We observed roughly the same trend in the regular bacterial biome at the species level and in the dark bacterial biome (S Figure 11, File S11). In contrast, the regular virome of each dataset was relatively unique with very low amounts of overlap (<= 3) between datasets (species and genus shows a similar pattern). Interestingly, the highest overlap for species in the dark virome was between our dataset and Ladd2017 (13), one of which is similar to RDRP viruses, although the contigs in Ladd's data were not similar to the velvet tobacco mottle virus in our dataset (Figure S12, File S12).

In addition to comparing datasets using taxonomy, we also compared contigs between datasets for nucleotide similarity (> 70%) using LAST (File S1 for methods). We found that in general, datasets in the regular biome with the largest amount of contigs have the most overlap. Unsurprisingly, in the dark biome we observed less overlap by nucleotide similarity and that our RDRP contig does not share nucleotide similarity with contigs from any dataset. In addition, we also compared the nucleotide similarity of double dark biome contigs and found there is not a large percentage of similar contigs between datasets (File S13).

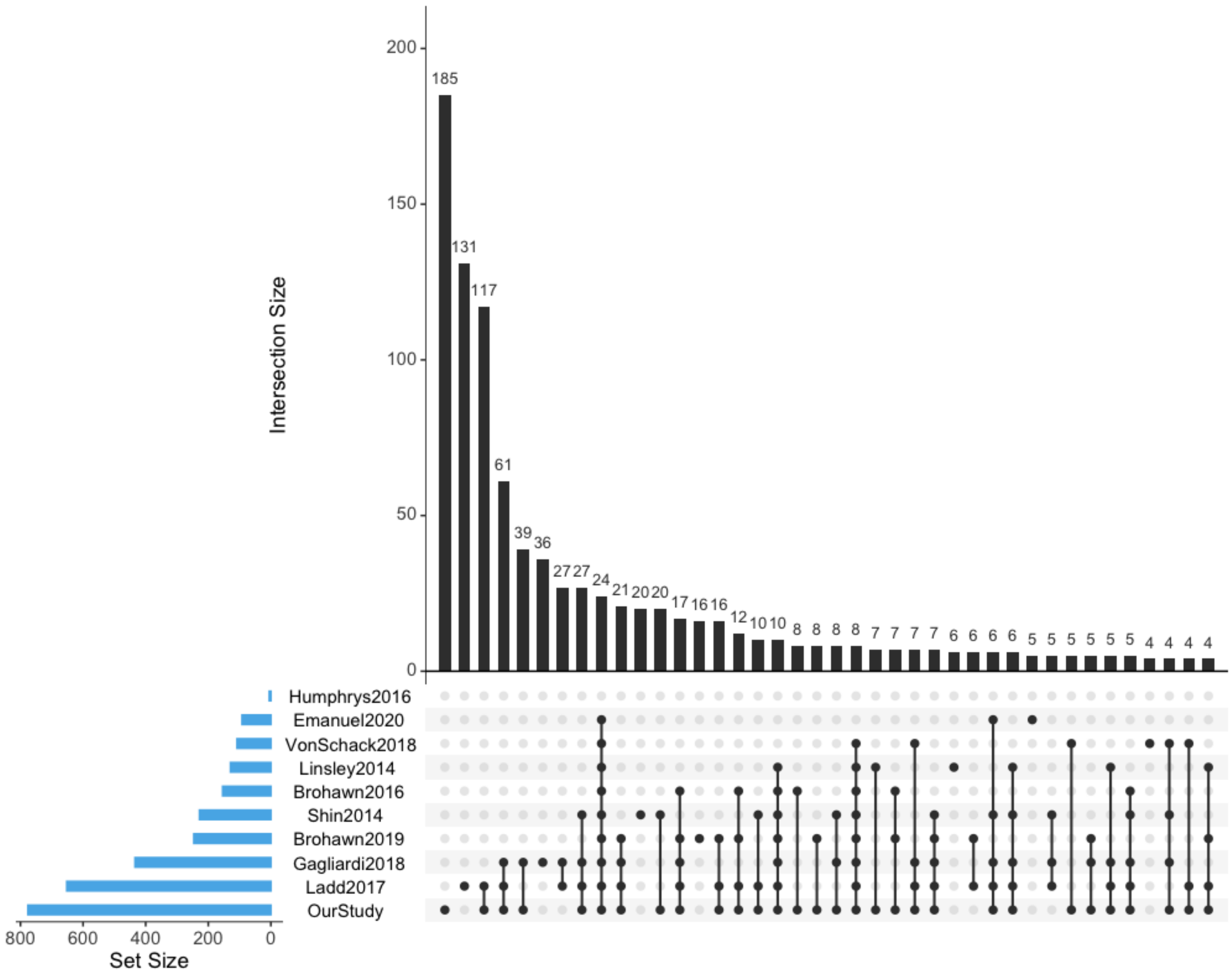

Figure 6. Upset plots of overlapping genus in the regular bacteriome between datasets. Upset plots are Venn diagram-like plots. A set refers to a dataset used in this study and each set is on a row with total amounts in a set as a blue bar plot on the left (ordered by set size). The black histogram on top shows the counts that are in the intersection of sets (a single dot for one dataset or connected dots for overlap of multiple datasets). Intersections less than 4 are removed for visualization purposes.

Comparison of taxonomy by condition within ALS datasets

Finally, we looked for differences in ALS vs control samples for each dataset. In the Gagliardi2014 dataset, when we compared ALS patients with the *FUS* mutation to controls, we found 3 significant differences in BNC by species in the regular bacteriome (*Neisseria sp.*, *Pseudomonas sp.*, *Sphingomonas sp.*) and one significant difference in BNC by genus in the dark bacteriome (*photobacterium)*. In ALS patients with mutations in *SOD1* compared to controls, we found two species significantly different in the regular bacteriome (*Hydrogenophaga crassostreae*, *Sphingomonas hengshuiensis) (Gagliardi FUS and SOD1 supplement)*. We did not find anything significant in sporadic ALS, or in ALS patients with *TARDBP* mutations with regards to genus/species (regular or dark biome or viruses) for Gagliardi2014. We found no significant statistical differences between ALS and control samples for genus/species of viruses/bacteria in the regular/dark biome for any of the remaining ALS datasets.

Meta analysis between datasets

Since our dataset and many others had no significant comparisons for ALS vs control groups within each dataset, a meta-analysis between datasets using this criteria would be difficult. As a second pass analysis we created a less stringent filtering method in order to compare the presence of microbes for each group between datasets (ALS vs. ALS; or controls vs. controls) (Figure 7). We assigned a contig to a condition if ≥ 2 samples from that condition contain at least 90% of the summed normalized coverage (from all samples) to the contig. This filtering greatly reduced the number of comparable genus/species for each dataset and, for example, reduced the genus of the regular bacteriome in our dataset from 305 for all samples to 33 (SYM:8, C9S:6, C9A:2, CTL:17) (File S14).

When we looked at ALS or control contigs in the regular bacteriome, the highest number of unique genus or species was from Ladd2017, and in general there was a small amount of overlap between datasets (≤1 for ALS or ≤ 8 for controls) (Figure 7). When we looked at genus in the dark bacteriome we saw no overlap for ALS contigs and low overlap (≤ 1) between control conditions (species was similar) (File S14). In the regular virome there was no overlap between datasets and only our study (one contig from ALS) and Ladd2017 (three from ALS, five from controls) had contigs that passed the filter (similar values for species). When we looked in the dark virome by genus there was no overlap between datasets, and our dataset had only one genus (*Sobemovirus* from controls*)* with the rest coming from Ladd2017 (18 from controls, 5 from ALS) (File S15). In conclusion, despite our conservative and loose approaches, we did not find any convincing evidence in ALS samples that the presence (or lack of presence) of an organism (or multiple organisms) was different compared to control samples.

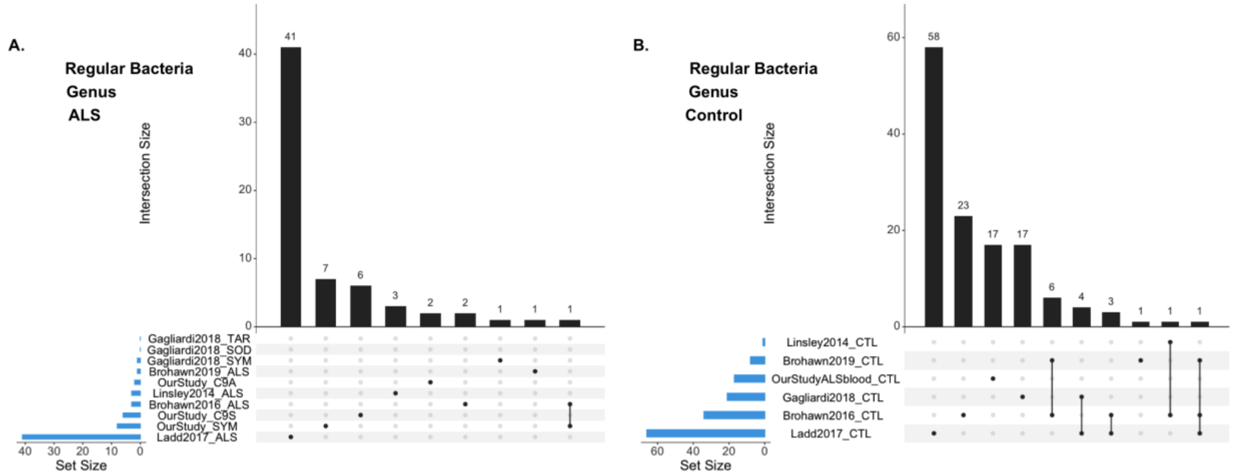

Fig 7. Upset plots of overlapping genus between datasets in the regular biome for ALS or controls.

Upset plots are Venn diagram-like plots. A set refers to a contig that was assigned to a condition from a dataset. Each set is on a row with total amounts in a set as a blue bar plot on the left (ordered by set size). The black histogram on top shows the counts that are in the intersection of sets (a single dot for one dataset or connected dots for overlap of multiple datasets). A. ALS contigs in the regular bacteriome. B. Control contigs from the regular bacteriome.

Discussion

We have created Mystery Miner to search for and quantify known and unknown microbes in RNA-seq datasets as a tool for researchers to study dysbiosis and identify infectious agents. We validated the pipeline by recovering and quantifying *Chlamydia* and SARS-CoV reads from RNA-seq datasets from intentionally infected cells. Interestingly, we also identified *Mycoplasma* reads in the *Chlamydia* dataset, suggesting that Mystery Miner may also be able to identify unsuspected bacterial infections. We also use published data to investigate the difference of polyA vs total RNA recovery of bacterial species in multiple tissues. Perhaps surprisingly, we did not see a consistent difference in the recovery of bacterial reads between the two types of RNA-seq libraries, considering that bacterial transcripts are not expected to be polyadenylated. However, it is well-recognized that polyA selection is imperfect, and libraries constructed from polyA-selected RNA routinely contain transcripts thought not to be polyadenylated (e.g., rRNA). We also found increased recovery of bacterial species with globin RNA depletion in blood. This could be the result of an effective increase in read depth for bacteria when not sequencing globin, or an increase in contamination from the globin depletion step. We stress that our bioinformatic approach alone cannot distinguish between contamination and the true existence of microbial sequences in human tissue.

We then used Mystery Miner on Our Study dataset consisting of 8.64 X 10^9 reads. This dataset was generated from whole blood total RNA that was depleted for both ribosomal and globin transcripts. It encompasses samples from controls, participants with a *C9ORF72* hexanucleotide expansion (symptomatic and pre-symptomatic), and *C9ORF72* negative ALS patients. We found no statistical difference in microbial sequence read coverage between controls and any class of ALS patients, examining either individual contigs or using various taxonomical binning approaches. We also did not detect any batch effects or obvious age- or sex- biases in the recovery of microbial reads (Figure S7). Overall, we found no evidence of blood microbial sequences contributing to either *C9ORF72* negative ALS or symptomatic patients harboring the *C9ORF72* hexanucleotide expansion. However, ALS is a CNS disease, so our findings in these blood samples do not necessarily preclude a role for microbes in this disease.

A unique aspect of Mystery Miner is that it tracks non-human reads that do not have significant BLASTN hits in GenBank. We were intrigued by the identification of a large contig (>5kb) in the dark biome of our ALS dataset that showed protein sequence similarity to RNA-dependent RNA polymerases, the essential replicase of RNA viruses. To validate that this virus-like sequence was not an artifact of contig assembly or a contaminant introduced during library construction or sequencing, we used RT-PCR of the original patient samples to demonstrate that this sequence was present in positive samples identified through the RNA-seq analysis and not detectable in negative samples. We cannot extrapolate from this specific example to determine what fraction of the "dark" and "double dark" sequences represent true novel microbial sequences present in human blood, but we note that analysis of human cell free blood DNA has reported the identification of thousands of novel bacterial sequences[86]. We suggest that Mystery Miner is a generally useful tool for the identification of novel microbial sequences in RNA-seq data.

To extend our analysis we processed publicly available blood and spinal cord ALS datasets through our pipeline. As observed in our dataset, library size generally correlated with number of bacterial contigs assembled and number of bacterial genera/species recovered. When the microbial sequences we found in our dataset were compared to the other datasets we found similar genera/species and, not surprisingly, larger datasets generally had greater overlap. One dataset (Ladd2017) yielded comparable recovery of bacteria and viruses for the regular biome but a far greater recovery bacteria and viruses in the dark biome compared to all the other datasets. This study used laser capture microdissection (LCM) to isolate cervical spinal cord motor neurons and had comparable read amounts per sample to other studies and was conducted in the same laboratory as two other studies (Brohawn2016, Brohawn2019). We are unsure why this dataset yielded a much larger dark biome compared to the other datasets. Potentially these differences are a byproduct of using LCM to acquire samples.

We then analyzed several publicly available ALS datasets for statistically significant differences between recovered microbial sequences in ALS and control samples. Only one dataset (Gagliardi2018) had any significant microbial sequence differences between control and ALS samples, specifically ALS patients with *FUS* or *SOD1* mutations. However, the sample number in this sub-study was small (N = 2-3), and in the case of the *SOD1* patients the excess microbial reads were in the control samples. In the absence of additional information (e.g., batch assignments for the samples) it is difficult to conclude that these sequence/sample correlations are meaningful. Finally, we compared identified microbial sequences in the control and ALS samples across the datasets and did not identify any genera/species that were preferentially recovered in either sample type.

Using our bioinformatic analysis pipeline Mystery Miner, we have not identified an association between ALS pathology and sequences corresponding to known or unknown microbial species. However, there are intrinsic limitations in using "repurposed" RNA-seq data to assay tissue-associated microbial sequences, including the relatively small number of non-human reads (<1% of total) upon which the analysis is based. This limited sequence signal could preclude identification of rarer microbes. Perhaps more problematic is the possibility that contaminating sequences obscure true tissue-associated microbial sequences. Any candidate microbes identified using Mystery Miner that correlate with human phenotypes will necessarily require independent validation. Despite these limitations, we believe Mystery Miner will be a useful tool for future researchers investigating known and unknown microbes that could contribute to disease, as our analyses have shown it to be sensitive to bacterial/viral agents in sequencing data.

Acknowledgements

We thank the Jackson Laboratories for globin depletion and RNA-seq to generate Our Study dataset. This work was supported by the NIH (NS063964, to CL); the ALS Association (ALSA, to LP; #WA1096 and 16-IIP-253 to MP; and #2015 to MB); NINDS (R35NS097273, P01NS084974, to LP); the Mayo Clinic Foundation (to LP); Neuroscience Focused Research Team Mayo Clinic grant (to MP); the Association of Frontotemporal Dementia (AFTD, to LP);

the Alzheimer's Association-AD Strategic Fund (to LP); the Muscular Dystrophy Association (MDA #172123 to MB); the ALS Recovery Fund (to MB); Kimmelman Estate (to MB); and the Department of Defense (Chem-Bio Diagnostics program Grant HDTRA-1-18-1-0032 to RDD).

Condition	Sample	GAPDH RT-PCR Ct Value	RDRP RT-PCR Ct Value	RDRP RNA-seq Normalized Coverage
SYM	LP00274	20.562019	36.401	1.56
C9S	LP00041	20.783213	36.346	3.39
C9S	LP00192	20.899612	35.636	0.67
C9A	LP000180	19.982108	34.832	1.11
C9S	LP000183	20.176418	undetermined	0
C9S	LP000197	20.125161	undetermined	0
C9A	LP000157	20.062433	undetermined	0

TABLE 1. RT-PCR AND NORMALIZED COVERAGE RESULTS FOR RDRP CONTIG
Quantitative RT-PCR and normalized coverage results from the 5180 bp RDRP contig. For the RDRP contig positive samples (top 4) with high normalized coverage and detectable Ct values and negative samples (bottom 3) with no normalized coverage and undetectable Ct values. GAPDH was used as a positive control for qRT-PCR and shows comparable levels for all samples. These samples were from three conditions *C9ORF72* negative ALS symptomatic patients (SYM), *C9ORF72* positive ALS symptomatic patients (C9S) and *C9ORF72* positive asymptomatic individuals (C9A).

Name	Groups	# Samples	Tissue	Pulldown
Humphrys2016	**1- or 24-hours post infection with *Chlamydia trachomatis***	**12**	**Cultured epithelial cell monolayers**	**PolyA** **Total RNA**
VonSchack2018	**PolyA or Total RNA from blood or colon**	**16**	**Whole Blood** **Colon**	**PolyA RNA** **Total RNA**
Shin2014	**Globin depleted** **Not globin depleted**	**24**	**Whole Blood**	**Total RNA**
Emanuel2020	**Severe acute respiratory syndrome coronavirus 1 or 2 infection** **Controls**	**18**	**Calu3 cells**	**Total RNA**
Our Study	***C9ORF72* negative ALS,** ***C9ORF72* positive and symptomatic ALS,** ***C9ORF72* positive asymptomatic participants** **Controls**	**120**	**Whole Blood**	**Total RNA hemoglobin and rRNA depleted**
Linsley2014	**ALS** **type 1 diabetes,** **sepsis,** **multiple sclerosis patients before and 24 hours after the first treatment with IFN-beta** **Controls**	**134**	**Whole blood**	**Total RNA**
Gagliardi2018	**Sporadic ALS,** **ALS with mutations in *FUS, SOD1, TARDBP*** **Controls**	**20**	**Peripheral blood mononuclear cells**	**Total RNA**
Brohawn2016	**ALS** **Controls**	**15**	**Cervical spinal cord**	**Total RNA** **rRNA depleted**

Ladd2017	**ALS** **Controls**	**10**	**Laser capture microdissection (LCM) to isolate cervical spinal cord motor neurons**	**Total RNA**
Brohawn2019	**ALS, Alzheimer's disease (AD), Parkinson's disease (PD)** **Controls**	**53**	**Cervical spinal cord**	**Total RNA**

TABLE 2. STUDY DESIGN FOR DATASETS USED IN THIS PAPER

Overview of the datasets used in this paper. The first three studies are only used to validate our pipeline. The six subsequent studies are ALS related from both blood and spinal cord.

Supplemental Figures

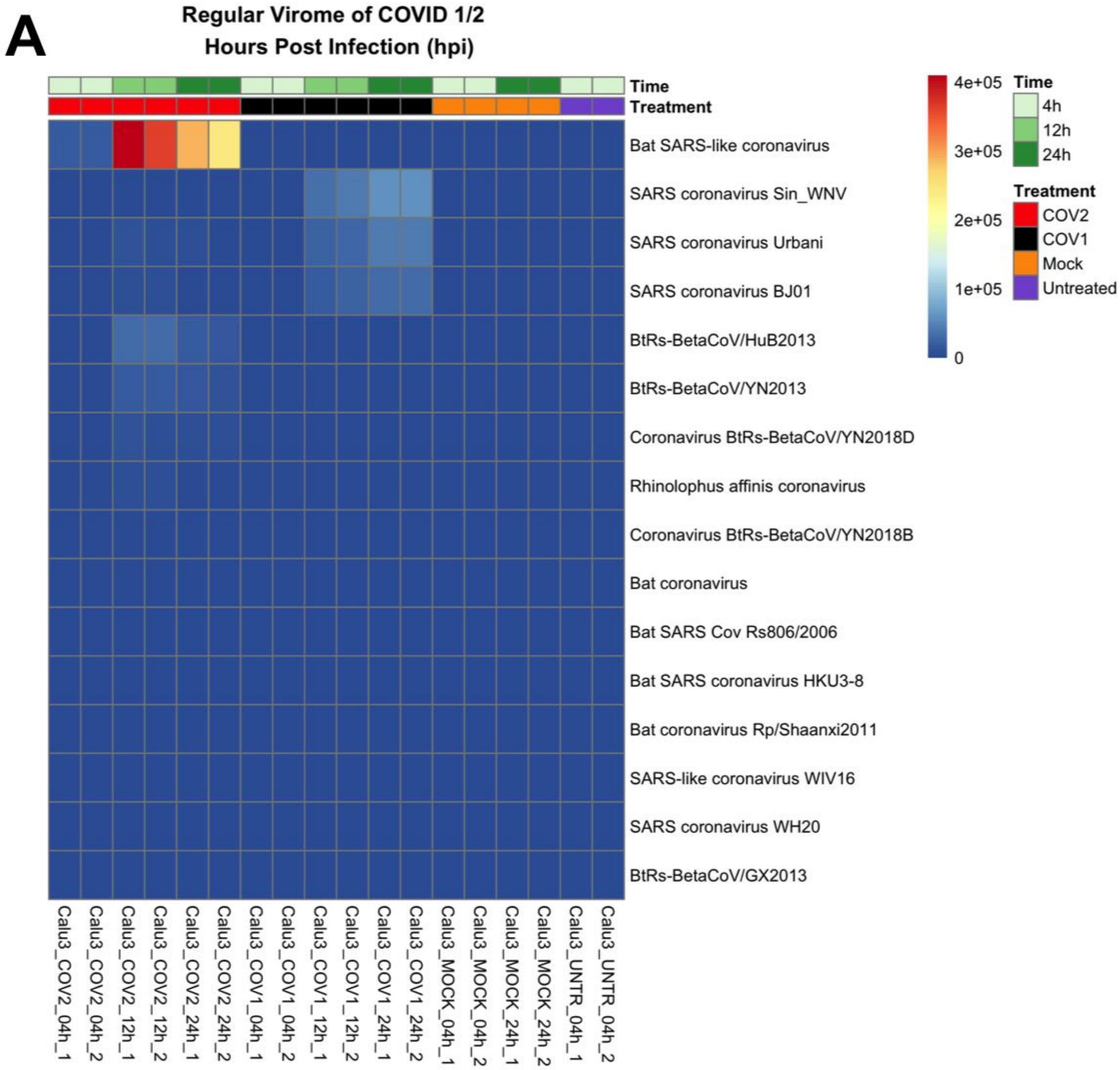

Figure S1. Heatmap of normalized coverage of regular Virome from Emanuel2020 with BLAST to nt database from 05/10/2019

Heatmap of normalized coverage of dark biome contigs binned by species (top 30 species). The nucleotide database was from 5/10/2019 before the discovery of SARS-CoV-2. The top row shows the same row from the main text but identified as a bat SARS like coronavirus.

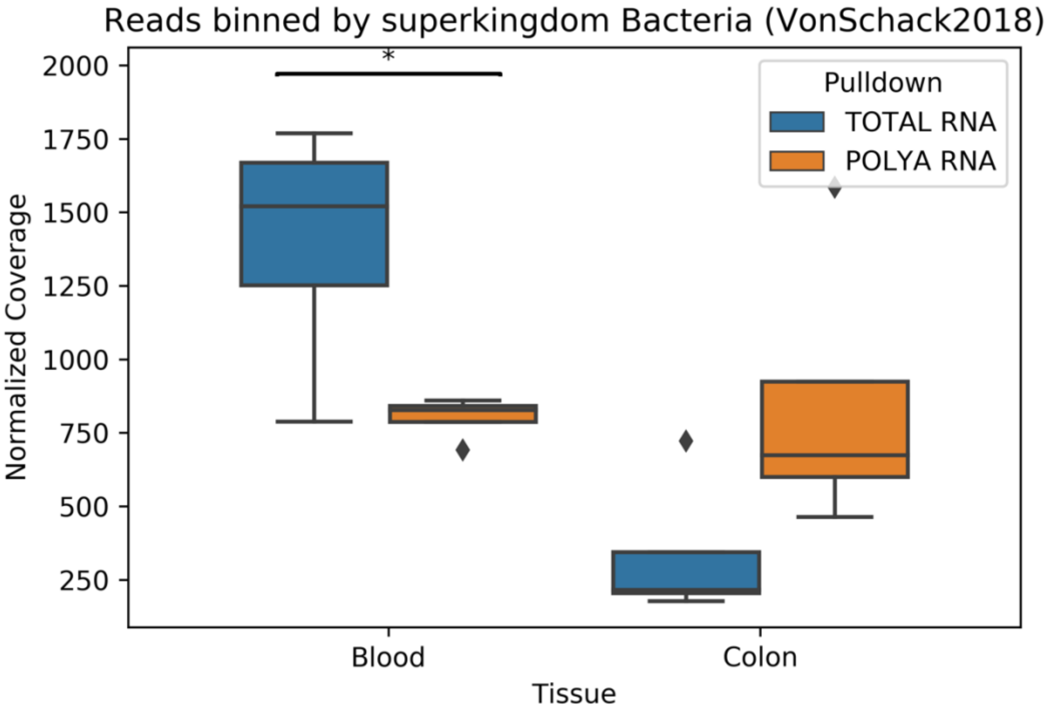

Figure S2. Boxplot of normalized coverage for superkingdom Bacteria in VonSchack2018 Boxplot of normalized coverage of regular biome contigs binned by superkingdom Bacteria. Blood shows significantly more reads in total RNA vs polyA RNA compared to Colon tissue.

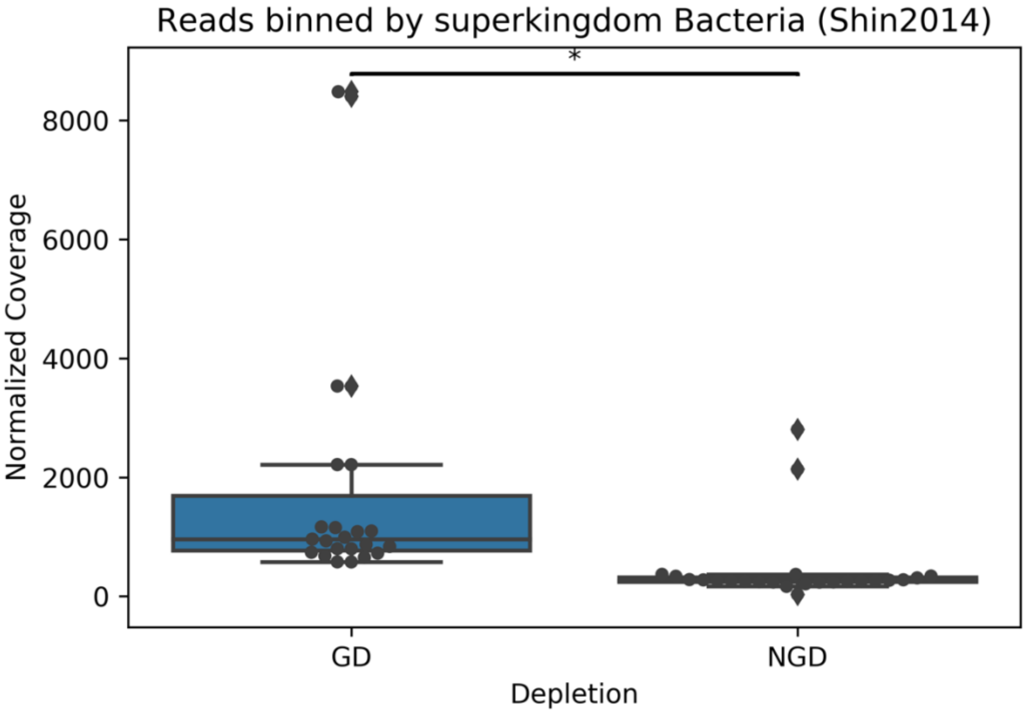

Figure S3. Boxplot of normalized coverage for superkingdom Bacteria in Shin2014
Boxplot of normalized coverage of regular biome contigs binned by superkingdom Bacteria. Globin depletion (GD) has significantly more coverage than non-globin depleted (NGD) blood.

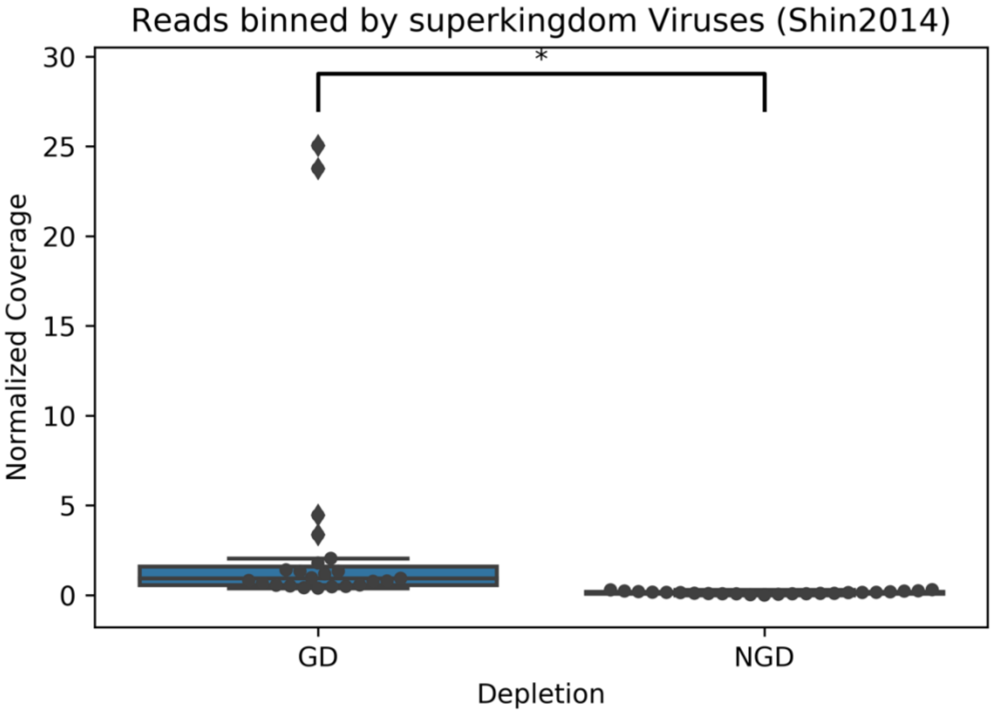

Figure S4. Boxplot of normalized coverage for superkingdom Viruses in Shin2014
Boxplot of normalized coverage of regular biome contigs binned by superkingdom Viruses. Globin depletion (GD) has significantly more coverage than non-globin depleted (NGD) blood.

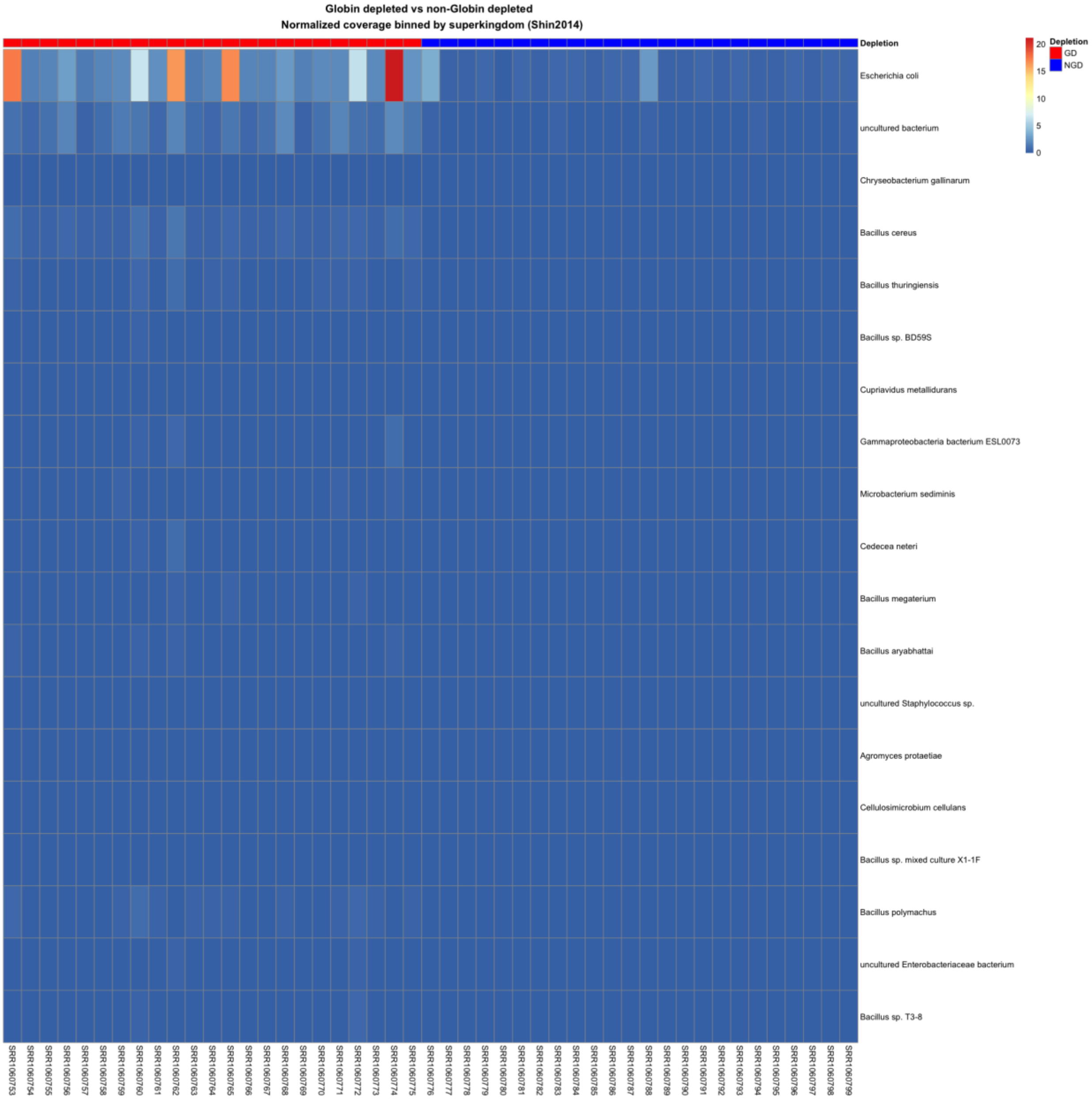

Figure S5. Heatmap of normalized coverage of regular Bacteriome binned by species from Shin2014

Heatmap of normalized coverage of regular biome contigs binned by bacteria species (top 20 species shown for brevity). Globin depletion (GD) is red and non-globin depletion is blue (NGD).

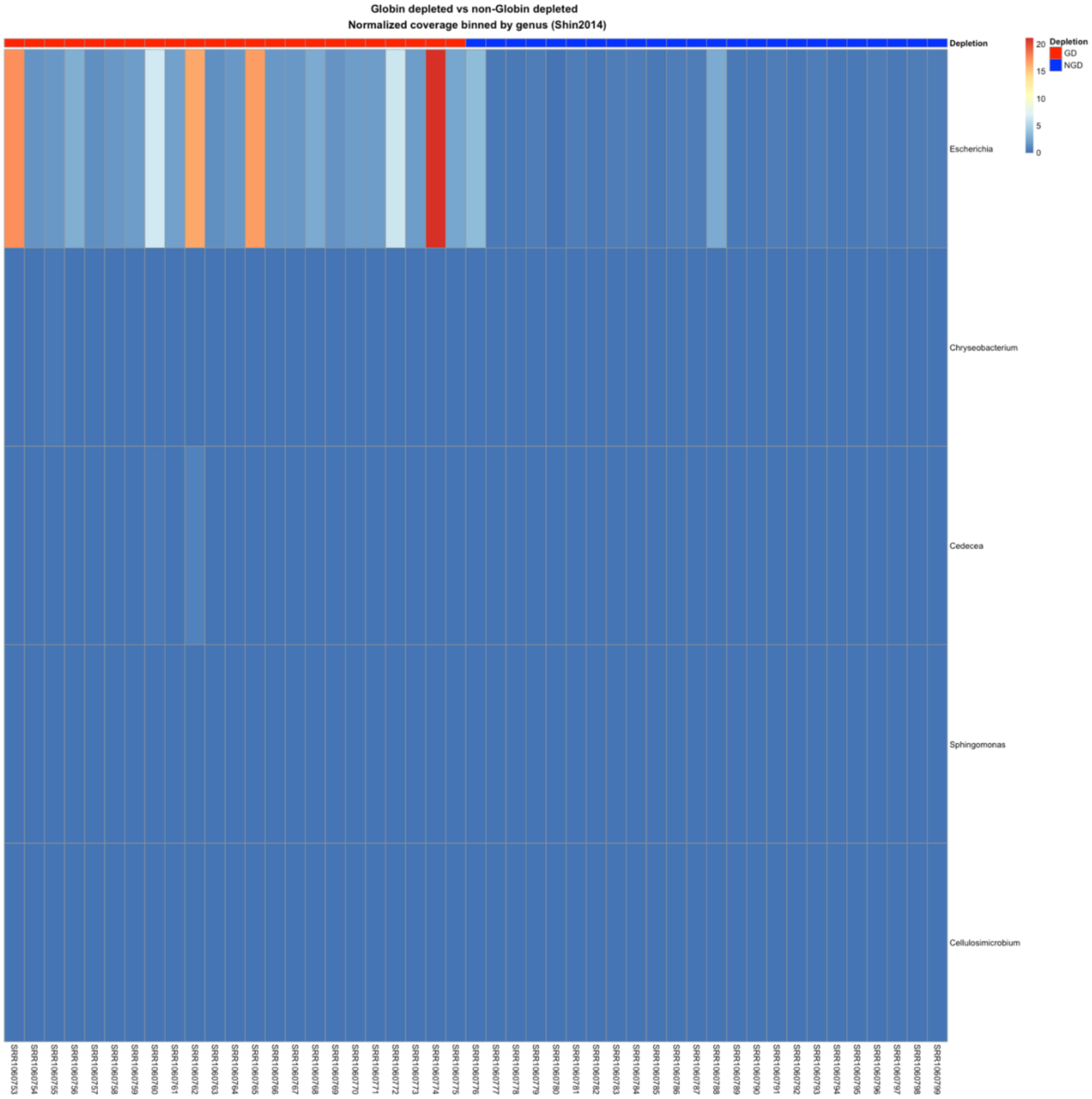

Figure S5. Heatmap of normalized coverage of regular Bacteriome binned by genus from Shin2014

Heatmap of normalized coverage of regular biome contigs binned by bacteria genus. Globin depletion (GD) is red and non-globin depletion is blue (NGD).

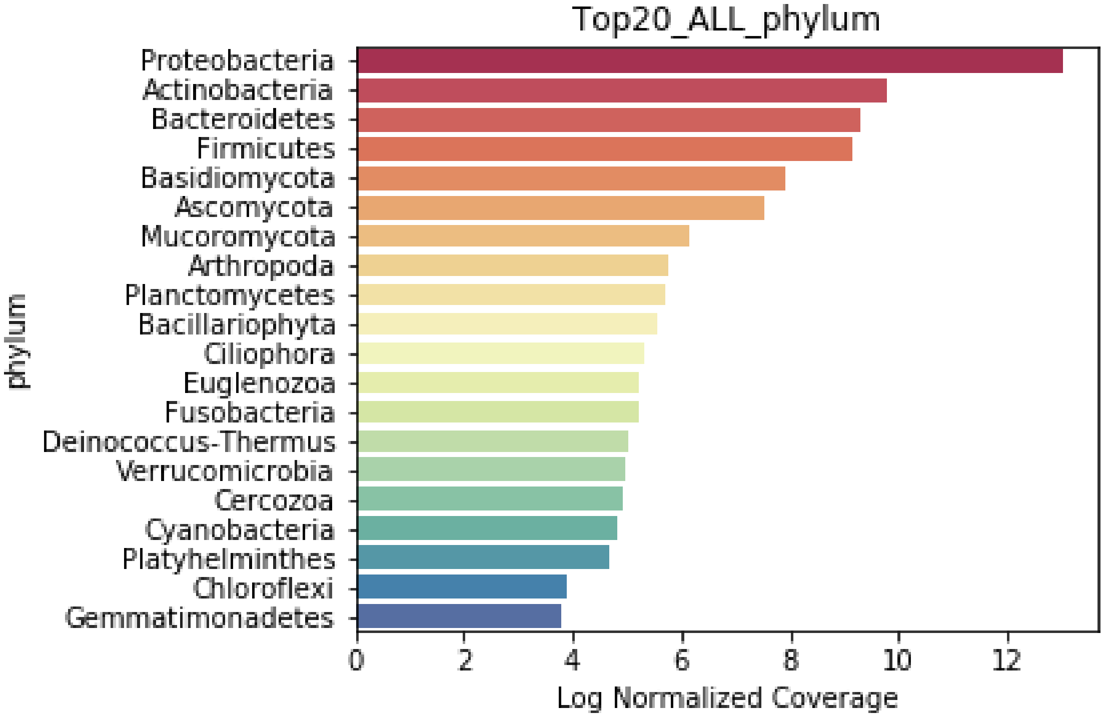

Figure S6. Log coverage binned by phylum from our ALS dataset
Coverage is summed for all of the samples and *alpha proteo-bacteria*, *Actinobacteria, Firmicutes,* and *Bacteroidetes* are the most highly represented.

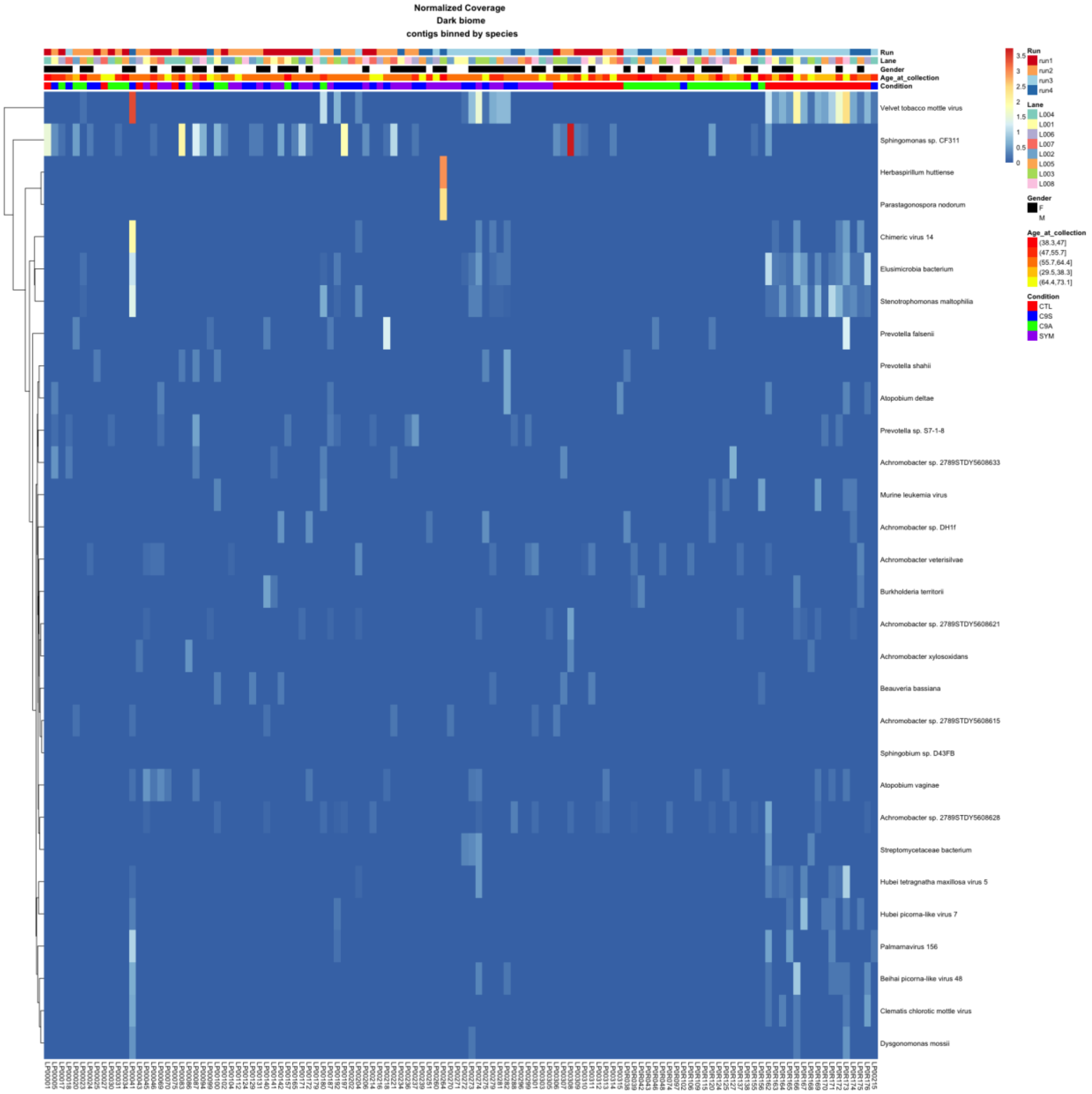

Figure S7. Heatmap of normalized coverage of dark biome contigs binned by species with metadata

Heatmap of normalized coverage of dark biome contigs binned by species (top 30 species shown for brevity). The highest coverage belongs to contigs that show high similarity to velvet tobacco mettle virus. Zero coverage is blue and goes to red with increasing values. These samples were from four conditions including control patients [(CTL) green], ALS symptomatic patients [(SYM) purple], C9-ORF positive ALS symptomatic patients [(C9S) blue] and C9-ORF positive asymptomatic patients [(C9A) red]. Other metadata include gender, lane, run, and age at collection.

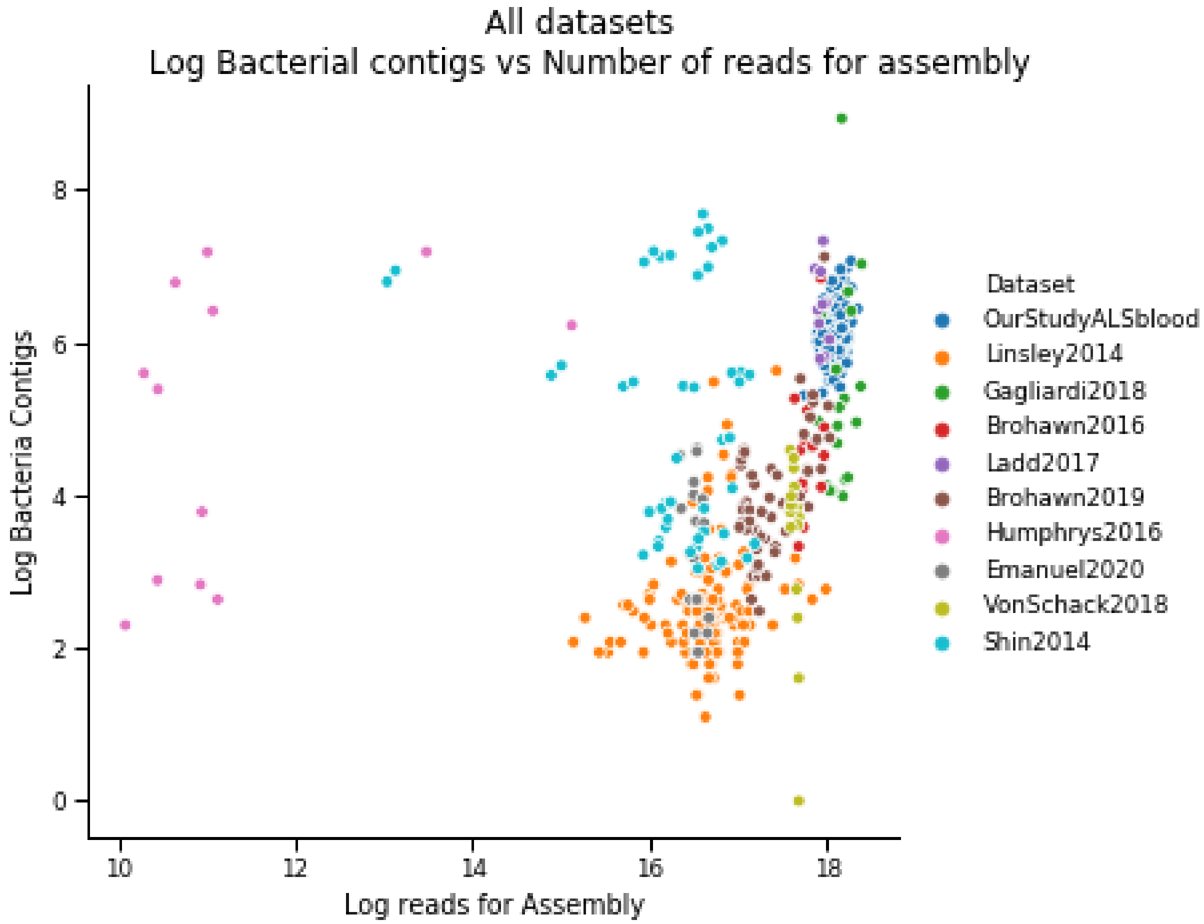

Figure S8. Log Bacterial contigs vs log reads for Assembly. Scatterplot where each dot is a sample from a dataset with log number of Bacterial contigs assembled on the Y-axis and Log reads used in SPAdes on the X-axis. Aside from the Shin, Humphrys, and Emanuel datasets there is a general trend of increased number of bacterial contigs with amount of reads.

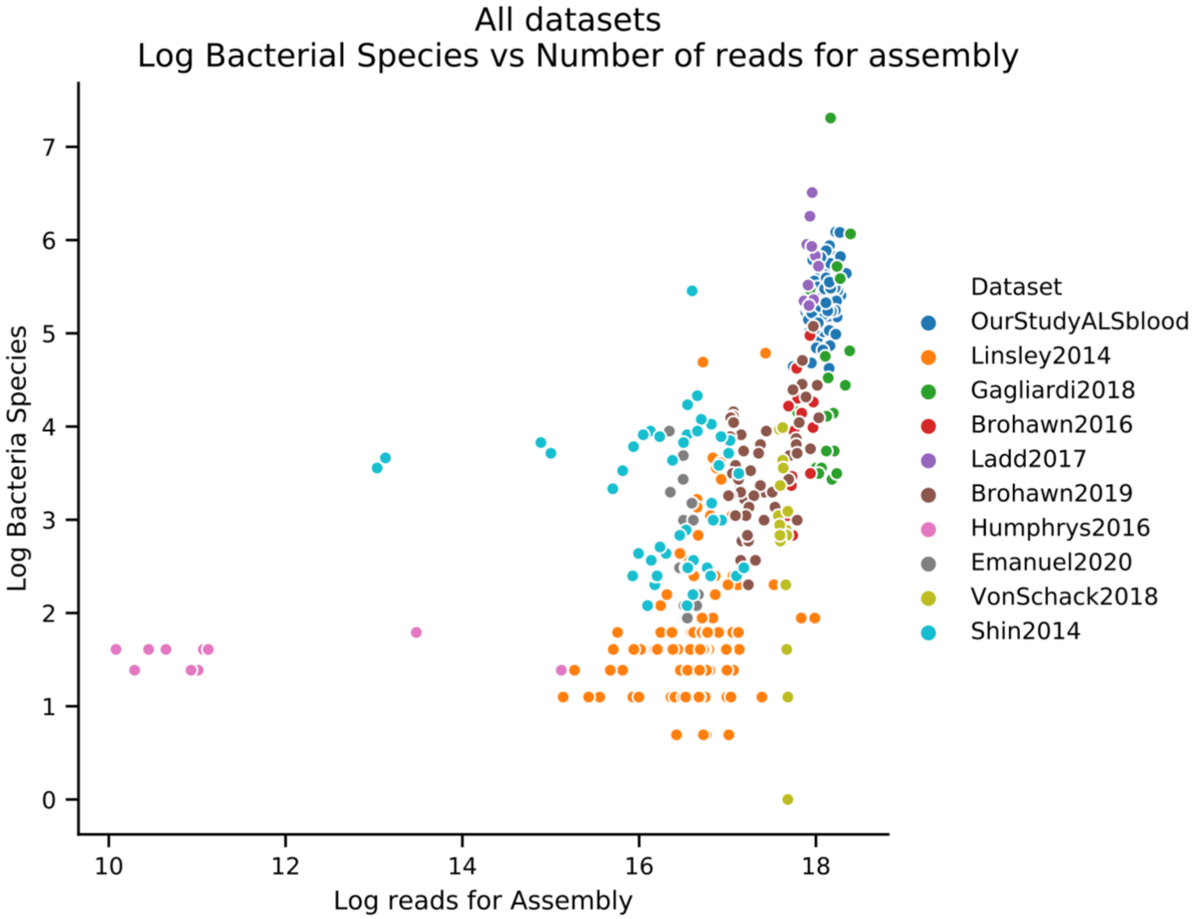

Figure S9. Log number of bacterial species vs log reads for Assembly. Scatterplot where each dot is a sample from a dataset with log number of number of bacterial species detected on the Y-axis and Log reads used in SPAdes on the X-axis.

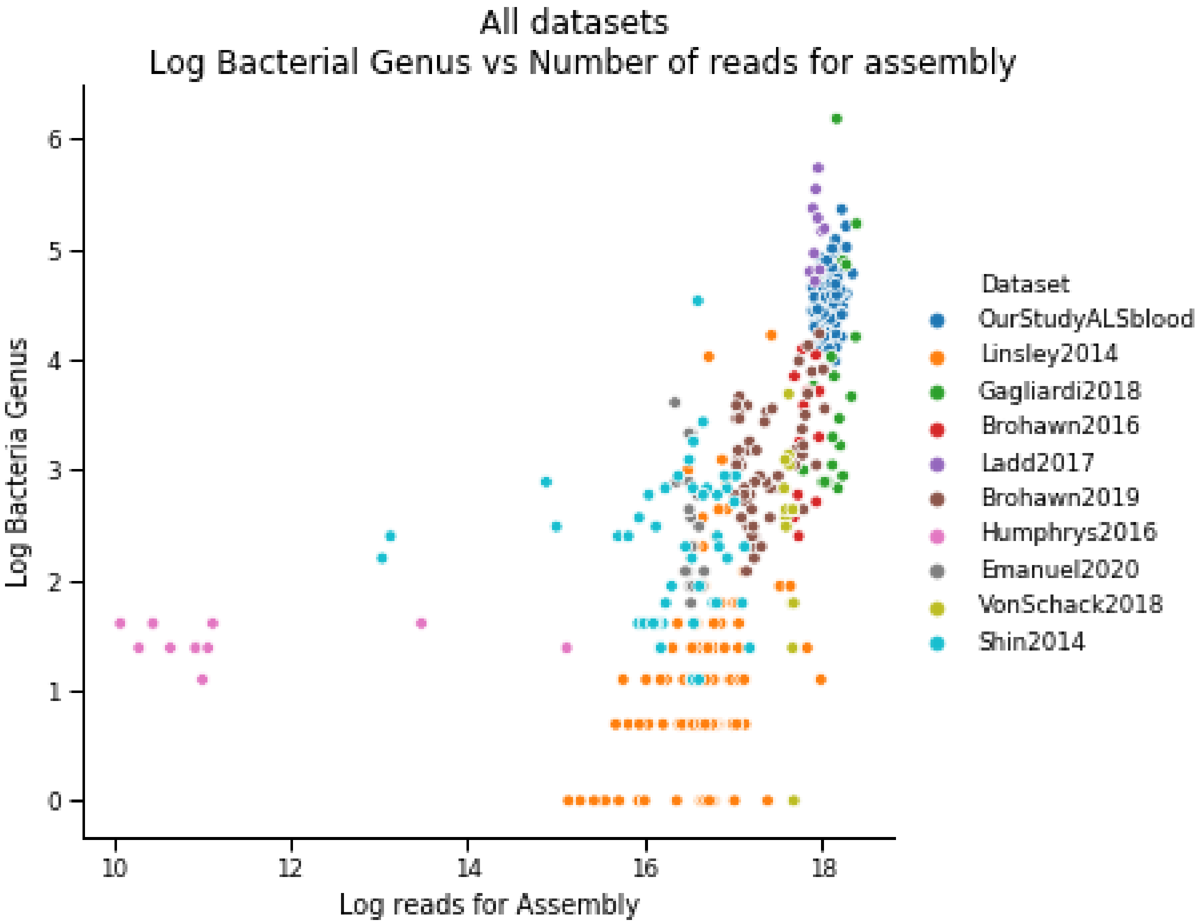

Figure S10. Log number of bacterial genus vs log reads for Assembly. Scatterplot where each dot is a sample from a dataset with log number of number of bacterial genus detected on the Y-axis and Log reads used in SPAdes on the X-axis.

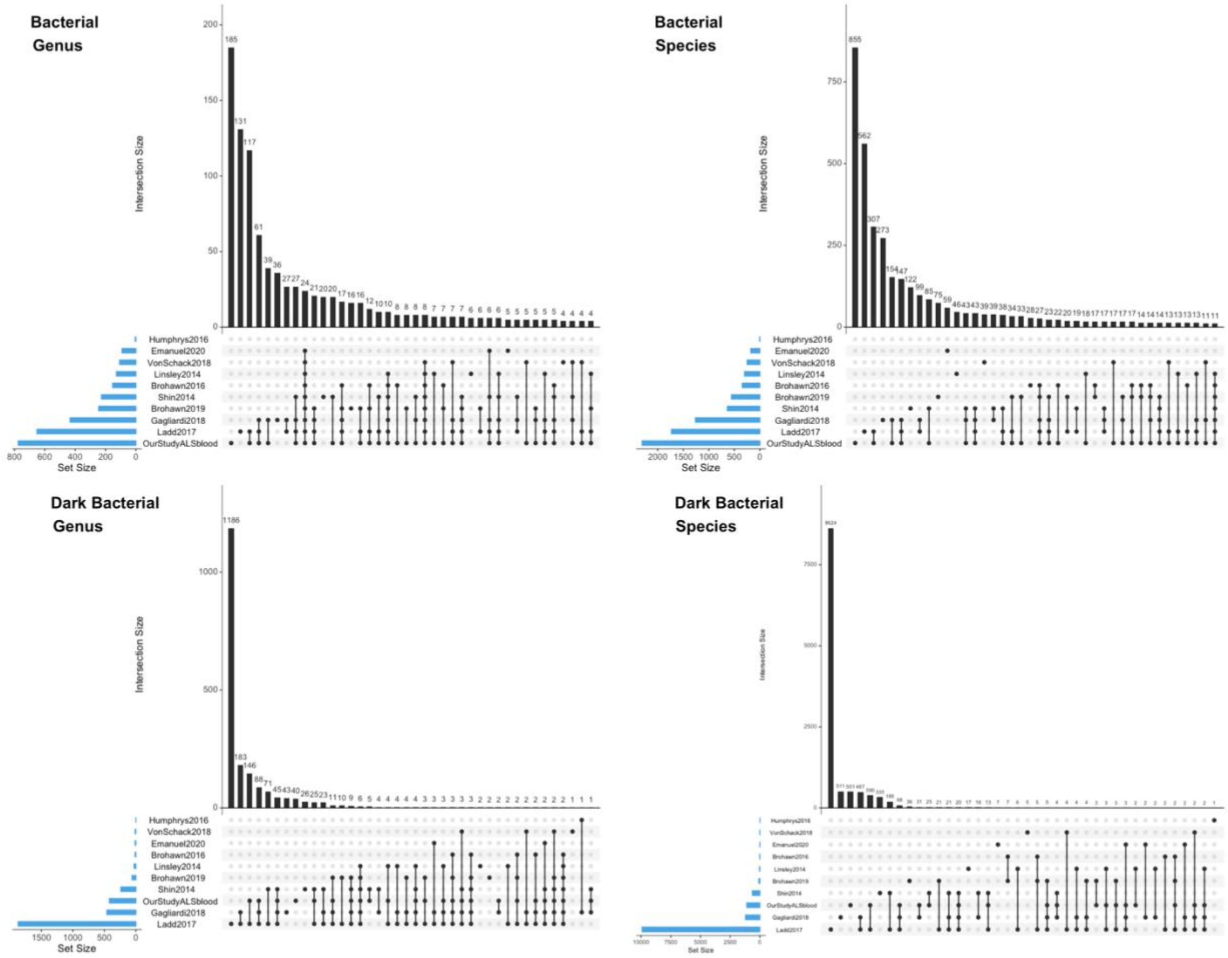

Figure S11. Upset plots of Bacteria for genus/species of regular/dark genome

Upset plots are venn diagram-like plots. Each set is on a row with total amounts in a set as a blue bar plot on the left. The black histogram on top shows the counts that are in the intersection of sets (a single dot for one set or connected dots for multiple sets). The highest number of overlapping bacterial genus is between our dataset and Ladd2017 (133) followed by the intersection between our dataset, Ladd2017 and Gagliardi2018 (61) and there is a modest overlap (24) for 9/10 datasets. This is roughly similar in the Bacterial species figure and in general the larger datasets have more unique and the highest number of overlap.

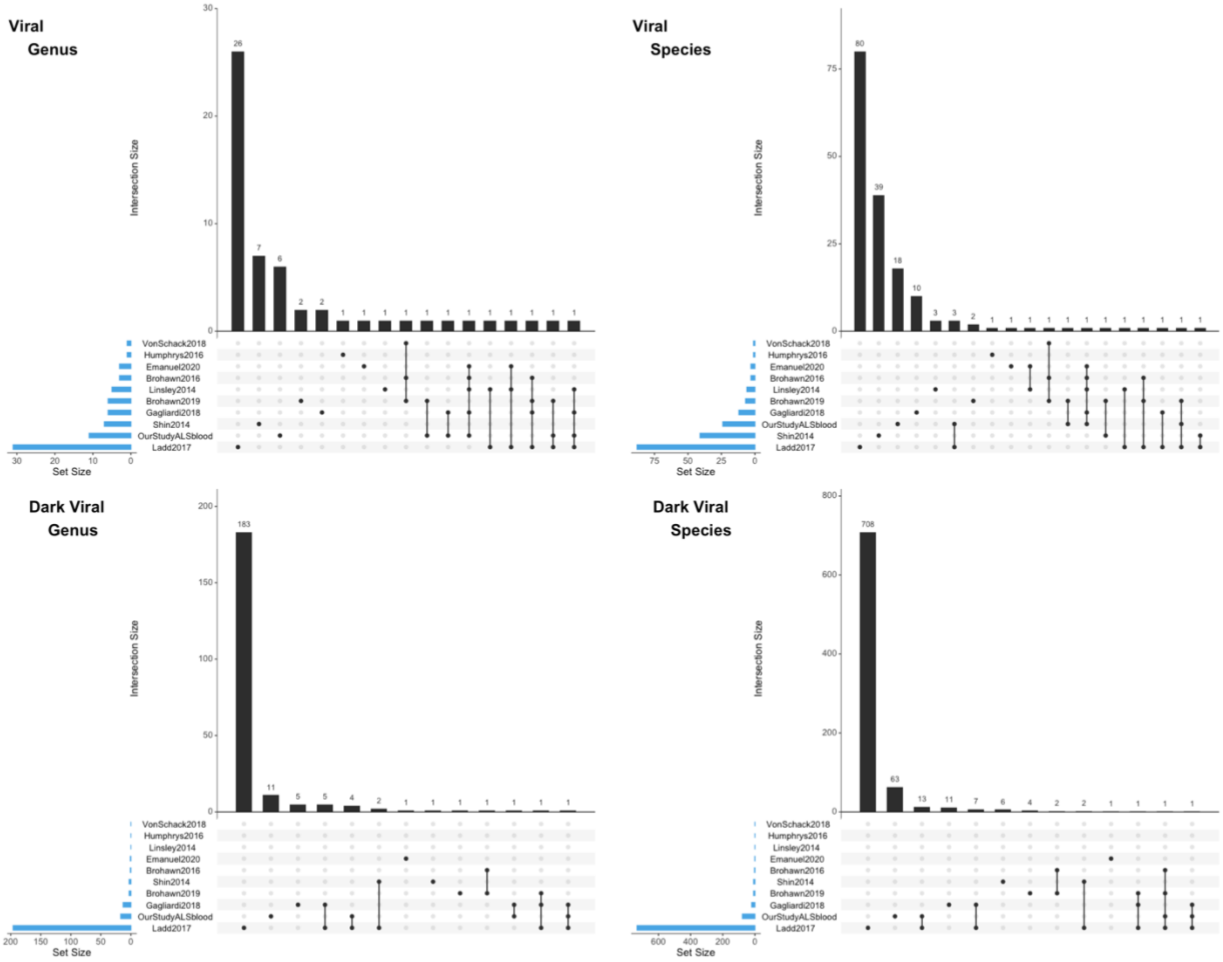

Figure S12. Upset plots of Viruses for genus/species of regular/dark genome
Upset plots are venn diagram-like plots. Each set is on a row with total amounts in a set as a blue bar plot on the left. The black histogram on top shows the counts that are in the intersection of sets (a single dot for one set or connected dots for multiple sets). The regular virome of each dataset is relatively unique with very low amounts of overlap (<= 3) between datasets (species and genus shows a similar pattern). Interestingly, the highest overlap for species in the dark virome is between our dataset and Ladd2017 (13).

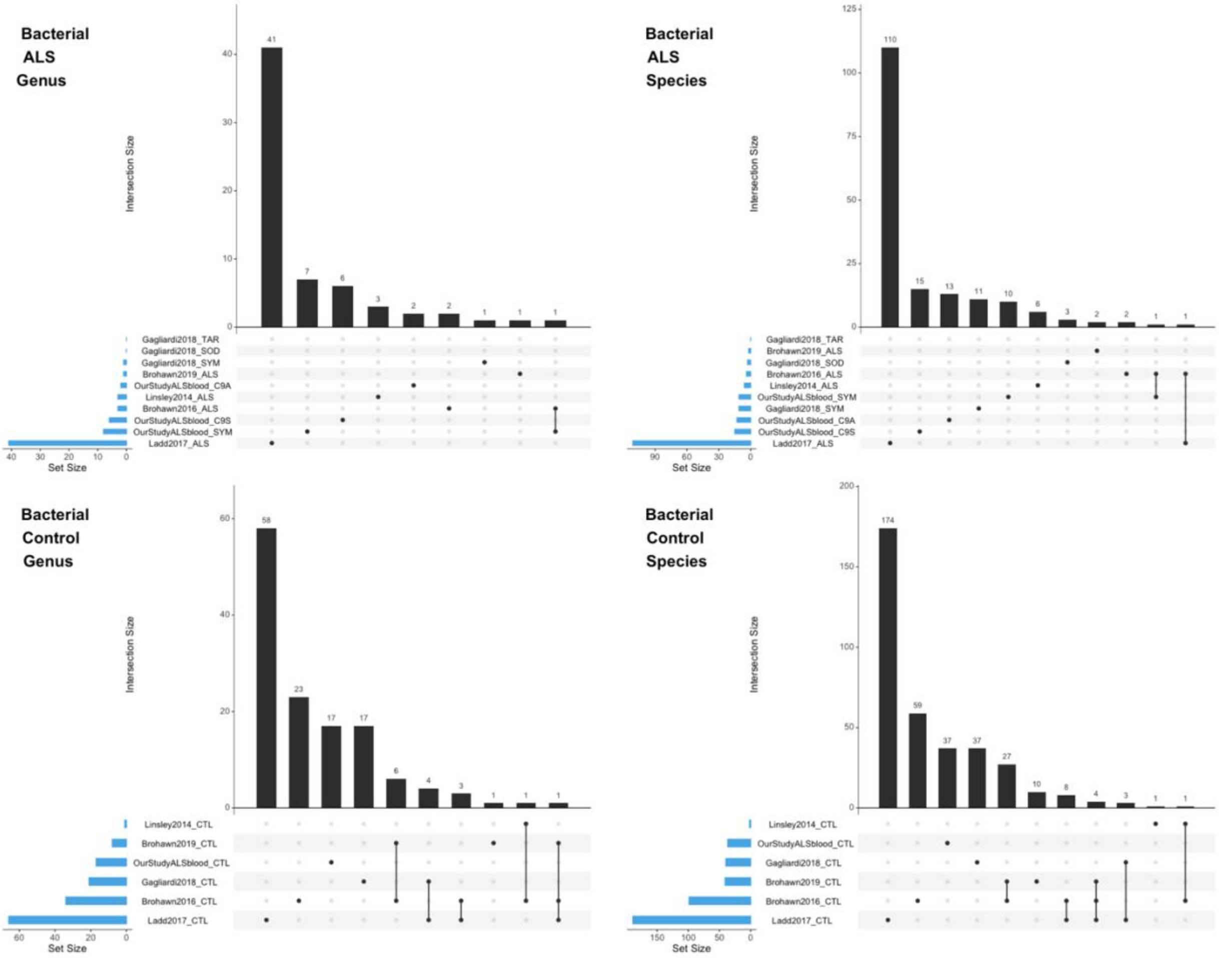

Figure S13. Upset plots of Bacteria in the regular biome for genus/species in ALS and Control contigs

Upset plots are venn diagram-like plots. Each set is on a row with total amounts in a set as a blue bar plot on the left. The black histogram on top shows the counts that are in the intersection of sets (a single dot for one set or connected dots for multiple sets). We assigned a contig to a condition if >= 2 samples from that condition contain at least 90% of the summed normalized coverage (from all samples) to the contig. In the genus and species from ALS samples there is a low amount of overlap between datasets (<= 1). When we look at control samples there is a much higher overlap for both genus and species.

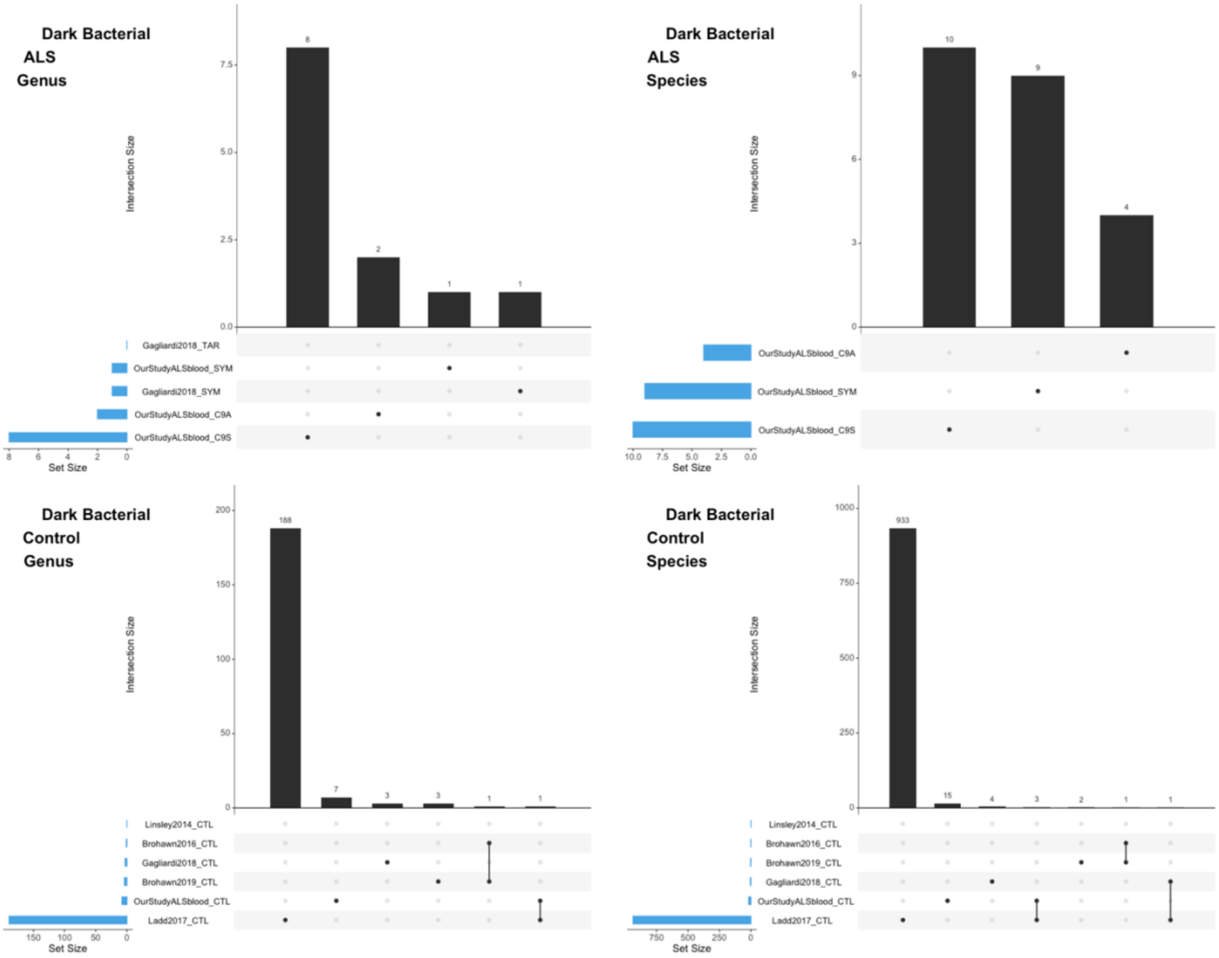

Figure S13. Upset plots of Bacteria in the dark biome for genus/species in ALS and Control contigs

Upset plots are venn diagram-like plots. Each set is on a row with total amounts in a set as a blue bar plot on the left. The black histogram on top shows the counts that are in the intersection of sets (a single dot for one set or connected dots for multiple sets). We assigned a contig to a condition if >= 2 samples from that condition contain at least 90% of the summed normalized coverage (from all samples) to the contig. Conditions with no recovered viruses have been omitted for clarity. Similarly to the regular bacteriome, there is no overlap in ALS samples and a small amount of overlap in the conditions.

References

1. Patrick KL, Bell SL, Weindel CG, Watson RO. Exploring the "multiple-hit hypothesis" of neurodegenerative disease: Bacterial infection comes up to bat. *Front Cell Infect Microbiol*. 2019. doi:10.3389/fcimb.2019.00138
2. Castanedo-Vazquez D, Bosque-Varela P, Sainz-Pelayo A, Riancho J. Infectious agents and amyotrophic lateral sclerosis: another piece of the puzzle of motor neuron degeneration. *J Neurol*. 2019. doi:10.1007/s00415-018-8919-3
3. Mehta P, Kaye W, Raymond J, et al. Prevalence of amyotrophic lateral sclerosis — United States, 2015. *Morb Mortal Wkly Rep*. 2018;67(46):1285-1289. doi:10.15585/mmwr.mm6746a1
4. Talbott EO, Malek AM, Lacomis D. The epidemiology of amyotrophic lateral sclerosis. In: *Handbook of Clinical Neurology*. Vol 138. Elsevier B.V.; 2016:225-238. doi:10.1016/B978-0-12-802973-2.00013-6
5. Ingre C, Roos PM, Piehl F, Kamel F, Fang F. Risk factors for amyotrophic lateral sclerosis. *Clin Epidemiol*. 2015. doi:10.2147/CLEP.S37505
6. Zhan Y, Fang F. Smoking and amyotrophic lateral sclerosis: A mendelian randomization study. *Ann Neurol*. 2019. doi:10.1002/ana.25443
7. Opie-Martin S, Wootton RE, Budu-Aggrey A, et al. Relationship between smoking and ALS: Mendelian randomisation interrogation of causality. *J Neurol Neurosurg Psychiatry*. 2020. doi:10.1136/jnnp-2020-323316
8. Kohne DE, Gibbs CJ, White L, Tracy SM, Meinke W, Smith RA. Virus detection by nucleic acid hybridization: Examination of normal and ALS tissues for the presence of poliovirus. *J Gen Virol*. 1981;56(2):223-233. doi:10.1099/0022-1317-56-2-223
9. Pertschuk LP, Broome JD, Brigati DJ, et al. JEJUNAL IMMUNOPATHOLOGY IN AMYOTROPHIC LATERAL SCLEROSIS AND MULTIPLE SCLEROSIS IDENTIFICATION OF VIRAL ANTIGENS BY IMMUNOFLUORESCENCE. *Lancet*. 1977;309(8022):1119-1123. doi:10.1016/S0140-6736(77)92382-0
10. Xue YC, Feuer R, Cashman N, Luo H. Enteroviral Infection: The Forgotten Link to Amyotrophic Lateral Sclerosis? *Front Mol Neurosci*. 2018;11:63. doi:10.3389/fnmol.2018.00063
11. Alonso R, Pisa D, Fernández-Fernández AM, Rábano A, Carrasco L. Fungal infection in neural tissue of patients with amyotrophic lateral sclerosis. *Neurobiol Dis*. 2017;108:249-260. doi:10.1016/j.nbd.2017.09.001
12. Andrade FC, Vergetti V, Cozza G, Falcao MC, Azevedo G. Amyotrophic Lateral Sclerosis-like Syndrome after Chikungunya. *Cureus*. October 2019. doi:10.7759/cureus.5876
13. Deutsch SI, Mohs RC, Davis KL. A rationale for studying the transmissibility of Alzheimer's disease. *Neurobiol Aging*. 1982;3(2):145-147. doi:10.1016/0197-4580(82)90011-2
14. Taylor GR, Crow TJ, Markakis DA, Lofthouse R, Neeley S, Carter GI. Herpes simplex virus and Alzheimer's disease: A search for virus DNA by spot hybridisation. *J Neurol Neurosurg Psychiatry*. 1984;47(10):1061-1065. doi:10.1136/jnnp.47.10.1061
15. Sochocka M, Zwolińska K, Leszek J. The Infectious Etiology of Alzheimer's Disease. *Curr Neuropharmacol*. 2017;15(7). doi:10.2174/1570159x15666170313122937

16. Irkeç C. [Virologic and immunologic considerations in Parkinson's disease]. *Mikrobiyol Bul*. 1982;16(4):293-296. http://www.ncbi.nlm.nih.gov/pubmed/6304477. Accessed December 9, 2019.

17. Abushouk AI, El-Husseny MWA, Magdy M, et al. Evidence for association between hepatitis C virus and Parkinson's disease. *Neurol Sci*. 2017;38(11):1913-1920. doi:10.1007/s10072-017-3077-4

18. Parashar A, Udayabanu M. Gut microbiota: Implications in Parkinson's disease. *Park Relat Disord*. 2017;38:1-7. doi:10.1016/j.parkreldis.2017.02.002

19. Libbey JE, Cusick MF, Fujinami RS. Role of pathogens in multiple sclerosis. *Int Rev Immunol*. 2014;33(4):266-283. doi:10.3109/08830185.2013.823422

20. Alonso R, Pisa D, Carrasco L. Searching for Bacteria in Neural Tissue From Amyotrophic Lateral Sclerosis. *Front Neurosci*. 2019;13:171. doi:10.3389/fnins.2019.00171

21. Gil C, González AAS, León IL, et al. Detection of Mycoplasmas in Patients with Amyotrophic Lateral Sclerosis. *Adv Microbiol*. 2014;04(11):712-719. doi:10.4236/aim.2014.411077

22. Alonso R, Pisa D, Marina AI, et al. Evidence for fungal infection in cerebrospinal fluid and brain tissue from patients with amyotrophic lateral sclerosis. *Int J Biol Sci*. 2015;11(5):546-558. doi:10.7150/ijbs.11084

23. Pisa D, Alonso R, Rábano A, Carrasco L. Corpora Amylacea of Brain Tissue from Neurodegenerative Diseases Are Stained with Specific Antifungal Antibodies. *Front Neurosci*. 2016;10:86. doi:10.3389/fnins.2016.00086

24. Cermelli C, Vinceti M, Beretti F, et al. Risk of sporadic amyotrophic lateral sclerosis associated with seropositivity for herpesviruses and echovirus-7. *Eur J Epidemiol*. 2003;18(2):123-127. doi:10.1023/a:1023067728557

25. Berger MM, Kopp N, Vital C, Redl B, Aymard M, Lina B. Detection and cellular localization of enterovirus RNA sequences in spinal cord of patients with ALS. *Neurology*. 2000;54(1):20-25. doi:10.1212/wnl.54.1.20

26. Vandenberghe N, Leveque N, Corcia P, et al. Cerebrospinal fluid detection of enterovirus genome in ALS: A study of 242 patients and 354 controls. *Amyotroph Lateral Scler*. 2010;11(3):277-282. doi:10.3109/17482960903262083

27. Xue YC, Feuer R, Cashman N, Luo H. Enteroviral infection: The forgotten link to amyotrophic lateral sclerosis? *Front Mol Neurosci*. 2018;11. doi:10.3389/fnmol.2018.00063

28. Giraud P, Beaulieux F, Ono S, Shimizu N, Chazot G, Lina B. Detection of enteroviral sequences from frozen spinal cord samples of Japanese ALS patients. *Neurology*. 2001;56(12):1777-1778. doi:10.1212/wnl.56.12.1777

29. Verma A, Berger JR. ALS syndrome in patients with HIV-1 infection. *J Neurol Sci*. 2006;240(1-2):59-64. doi:10.1016/j.jns.2005.09.005

30. Moodley K, Bill PLA, Bhigjee AI, Patel VB. A comparative study of motor neuron disease in HIV-infected and HIV-uninfected patients. *J Neurol Sci*. 2019;397:96-102. doi:10.1016/J.JNS.2018.12.030

31. Douville R, Liu J, Rothstein J, Nath A. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Ann Neurol*. 2011;69(1):141-151. doi:10.1002/ana.22149

32. Li W, Lee M-H, Henderson L, et al. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med*. 2015;7(307):307ra153-307ra153. doi:10.1126/scitranslmed.aac8201

33. Arru G, Mameli G, Deiana GA, et al. Humoral immunity response to human endogenous retroviruses K/W differentiates between amyotrophic lateral sclerosis and other neurological diseases. *Eur J Neurol*. 2018;25(8):1076-e84. doi:10.1111/ene.13648

34. Blacher E, Bashiardes S, Shapiro H, et al. Potential roles of gut microbiome and metabolites in modulating ALS in mice. *Nature*. 2019;572(7770):474-480. doi:10.1038/s41586-019-1443-5

35. Fang X, Wang X, Yang S, et al. Evaluation of the microbial diversity in amyotrophic lateral sclerosis using high-throughput sequencing. *Front Microbiol*. 2016;7(SEP). doi:10.3389/fmicb.2016.01479

36. Sun J, Zhan Y, Mariosa D, et al. Antibiotics use and risk of amyotrophic lateral sclerosis in Sweden. *Eur J Neurol*. 2019;26(11):1355-1361. doi:10.1111/ene.13986

37. Zhang Y guo, Wu S, Yi J, et al. Target Intestinal Microbiota to Alleviate Disease Progression in Amyotrophic Lateral Sclerosis. *Clin Ther*. 2017;39(2):322-336. doi:10.1016/j.clinthera.2016.12.014

38. Obrenovich M, Jaworski H, Tadimalla T, et al. The role of the microbiota–gut–brain axis and antibiotics in ALS and neurodegenerative diseases. *Microorganisms*. 2020;8(5). doi:10.3390/microorganisms8050784

39. Brenner D, Hiergeist A, Adis C, et al. The fecal microbiome of ALS patients. *Neurobiol Aging*. 2018;61:132-137. doi:10.1016/j.neurobiolaging.2017.09.023

40. Henkel JS, Beers DR, Wen S, Bowser R, Appel SH. Decreased mRNA expression of tight junction proteins in lumbar spinal cords of patients with ALS. *Neurology*. 2009;72(18):1614-1616. doi:10.1212/WNL.0b013e3181a41228

41. Garbuzova-Davis S, Sanberg PR. Blood-CNS barrier impairment in ALS patients versus an animal model. *Front Cell Neurosci*. 2014;8(FEB):21. doi:10.3389/fncel.2014.00021

42. Zhang R, Miller RG, Gascon R, et al. Circulating endotoxin and systemic immune activation in sporadic amyotrophic lateral sclerosis (sALS). *J Neuroimmunol*. 2009;206(1-2):121-124. doi:10.1016/j.jneuroim.2008.09.017

43. Mantovani S, Garbelli S, Pasini A, et al. Immune system alterations in sporadic amyotrophic lateral sclerosis patients suggest an ongoing neuroinflammatory process. *J Neuroimmunol*. 2009;210(1-2):73-79. doi:10.1016/j.jneuroim.2009.02.012

44. Murdock BJ, Zhou T, Kashlan SR, Little RJ, Goutman SA, Feldman EL. Correlation of peripheral immunity with rapid Amyotrophic lateral sclerosis progression. *JAMA Neurol*. 2017;74(12):1446-1454. doi:10.1001/jamaneurol.2017.2255

45. Sta M, Sylva-Steenland RMR, Casula M, et al. Innate and adaptive immunity in amyotrophic lateral sclerosis: Evidence of complement activation. *Neurobiol Dis*. 2011;42(3):211-220. doi:10.1016/j.nbd.2011.01.002

46. Correia AS, Patel P, Dutta K, Julien JP. Inflammation induces TDP-43 mislocalization and aggregation. *PLoS One*. 2015;10(10). doi:10.1371/journal.pone.0140248

47. Verber NS, Shepheard SR, Sassani M, et al. Biomarkers in motor neuron disease: A state of the art review. *Front Neurol*. 2019;10(APR):291. doi:10.3389/fneur.2019.00291

48. Blasco H, Corcia P, Moreau C, et al. 1H-NMR-Based metabolomic profiling of CSF in early

amyotrophic lateral sclerosis. *PLoS One*. 2010;5(10). doi:10.1371/journal.pone.0013223

49. Blasco H, Veyrat-Durebex C, Bocca C, et al. Lipidomics Reveals Cerebrospinal-Fluid Signatures of ALS. *Sci Rep*. 2017;7(1). doi:10.1038/s41598-017-17389-9
50. Mitchell RM, Freeman WM, Randazzo WT, et al. A CSF biomarker panel for identification of patients with amyotrophic lateral sclerosis. *Neurology*. 2009;72(1):14-19. doi:10.1212/01.wnl.0000333251.36681.a5
51. Guo J, Yang X, Gao L, Zang D. Evaluating the levels of CSF and serum factors in ALS. *Brain Behav*. 2017;7(3). doi:10.1002/brb3.637
52. Young PE, Jew SK, Buckland ME, Pamphlett R, Suter CM. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. *PLoS One*. 2017;12(8). doi:10.1371/journal.pone.0182638
53. Coppedè F, Stoccoro A, Mosca L, et al. Increase in DNA methylation in patients with amyotrophic lateral sclerosis carriers of not fully penetrant SOD1 mutations. *Amyotroph Lateral Scler Front Degener*. 2018;19(1-2):93-101. doi:10.1080/21678421.2017.1367401
54. Swindell WR, Kruse CPS, List EO, Berryman DE, Kopchick JJ. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J Transl Med*. 2019;17(1):170. doi:10.1186/s12967-019-1909-0
55. Waller R, Wyles M, Heath PR, et al. Small RNA sequencing of sporadic amyotrophic lateral sclerosis cerebrospinal fluid reveals differentially expressed miRNAs related to neural and glial activity. *Front Neurosci*. 2018;11(JAN). doi:10.3389/fnins.2017.00731
56. Waller R, Goodall EF, Milo M, et al. Serum miRNAs miR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS). *Neurobiol Aging*. 2017;55:123-131. doi:10.1016/j.neurobiolaging.2017.03.027
57. Gendron TF, Chew J, Stankowski JN, et al. Poly(GP) proteins are a useful pharmacodynamic marker for C9ORF72-associated amyotrophic lateral sclerosis. *Sci Transl Med*. 2017;9(383):7866. doi:10.1126/scitranslmed.aai7866
58. Gagliardi S, Zucca S, Pandini C, et al. Long non-coding and coding RNAs characterization in Peripheral Blood Mononuclear Cells and Spinal Cord from Amyotrophic Lateral Sclerosis patients. *Sci Rep*. 2018;8(1):2378. doi:10.1038/s41598-018-20679-5
59. Zucca S, Gagliardi S, Pandini C, et al. RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls. *Sci Data*. 2019;6(1):190006. doi:10.1038/sdata.2019.6
60. Rahman MR, Islam T, Huq F, Quinn JMW, Moni MA. Identification of molecular signatures and pathways common to blood cells and brain tissue of amyotrophic lateral sclerosis patients. *Informatics Med Unlocked*. 2019;16:100193. doi:10.1016/J.IMU.2019.100193
61. van Rheenen W, Diekstra FP, Harschnitz O, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS One*. 2018;13(6):e0198874. doi:10.1371/journal.pone.0198874
62. Parker J, Chen J. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. *J Clin Virol*. 2017;86:20-26. doi:10.1016/j.jcv.2016.11.010
63. Bouquet J, Gardy JL, Brown S, et al. RNA-Seq Analysis of Gene Expression, Viral Pathogen, and B-Cell/T-Cell Receptor Signatures in Complex Chronic Disease. *Clin Infect Dis*. 2017;64(4):476-481. doi:10.1093/cid/ciw767

64. Westermann AJ, Barquist L, Vogel J. Resolving host-pathogen interactions by dual RNA-seq. *PLoS Pathog*. 2017;13(2):e1006033. doi:10.1371/journal.ppat.1006033

65. Moore RA, Warren RL, Freeman JD, et al. The Sensitivity of Massively Parallel Sequencing for Detecting Candidate Infectious Agents Associated with Human Tissue. Jordan IK, ed. *PLoS One*. 2011;6(5):e19838. doi:10.1371/journal.pone.0019838

66. Poussin C, Sierro N, Boué S, et al. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. *Drug Discov Today*. 2018;23(9):1644-1657. doi:10.1016/j.drudis.2018.06.005

67. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front Genet*. 2017;8(MAR). doi:10.3389/fgene.2017.00023

68. Mangul S, Yang HT, Strauli N, et al. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol*. 2018;19(1):36. doi:10.1186/s13059-018-1403-7

69. Cavadas B, Ferreira J, Camacho R, Fonseca NA, Pereira L. QmihR: Pipeline for Quantification of Microbiome in Human RNA-seq. In: Springer, Cham; 2017:173-179. doi:10.1007/978-3-319-60816-7_21

70. Simon LM, Karg S, Westermann AJ, et al. MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. *Gigascience*. 2018;7(6). doi:10.1093/gigascience/giy070

71. Gihawi A, Rallapalli G, Hurst R, Cooper CS, Leggett RM, Brewer DS. SEPATH: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome Biol*. 2019;20(1):208. doi:10.1186/s13059-019-1819-8

72. Cox JW, Ballweg RA, Taft DH, Velayutham P, Haslam DB, Porollo A. A fast and robust protocol for metataxonomic analysis using RNAseq data. *Microbiome*. 2017;5(1):7. doi:10.1186/s40168-016-0219-5

73. Almeida A, Mitchell AL, Boland M, et al. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568(7753):499-504. doi:10.1038/s41586-019-0965-1

74. Papudeshi B, Haggerty JM, Doane M, et al. Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics*. 2017;18(1):915. doi:10.1186/s12864-017-4294-1

75. Humphrys MS, Creasy T, Sun Y, et al. Simultaneous transcriptional profiling of bacteria and their host cells. Ramsey K, ed. *PLoS One*. 2013;8(12):e80597. doi:10.1371/journal.pone.0080597

76. Emanuel W, Kirstin M, Vedran F, et al. Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv*. May 2020:2020.05.05.079194. doi:10.1101/2020.05.05.079194

77. Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol*. July 2020:1-10. doi:10.1038/s41564-020-0771-4

78. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep*. 2018;8(1):4781. doi:10.1038/s41598-018-23226-4

79. Shin H, Shannon CP, Fishbane N, et al. Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion. Wang K, ed. *PLoS One*. 2014;9(3):e91041. doi:10.1371/journal.pone.0091041

80. Castillo DJ, Rifkin RF, Cowan DA, Potgieter M. The healthy human blood microbiome: Fact or fiction? *Front Cell Infect Microbiol*. 2019;9(MAY):148. doi:10.3389/fcimb.2019.00148

81. Mackay IM, Arden KE, Nitsche A. Real-time PCR in virology. *Nucleic Acids Res*. 2002;30(6):1292-1305. doi:10.1093/nar/30.6.1292

82. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One*. 2014;9(10). doi:10.1371/journal.pone.0109760

83. Brohawn DG, O'Brien LC, Bennett JP. RNAseq analyses identify tumor necrosis factor-mediated inflammation as a major abnormality in ALS spinal cord. *PLoS One*. 2016;11(8):e0160520. doi:10.1371/journal.pone.0160520

84. C Ladd A, G Brohawn D, P Bennett J. Laser-captured spinal cord motorneurons from ALS subjects show increased gene expression in vacuolar ATPase networks. *J Syst Integr Neurosci*. 2017;3(6). doi:10.15761/jsin.1000182

85. Bennett JP, Keeney PM, Brohawn DG. RNA Sequencing Reveals Small and Variable Contributions of Infectious Agents to Transcriptomes of Postmortem Nervous Tissues From Amyotrophic Lateral Sclerosis, Alzheimer's Disease and Parkinson's Disease Subjects, and Increased Expression of Genes From Disease-Activated Microglia. *Front Neurosci*. 2019;13. doi:10.3389/fnins.2019.00235

86. Kowarsky M, Camunas-Soler J, Kertesz M, et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc Natl Acad Sci U S A*. 2017;114(36):9623-9628. doi:10.1073/pnas.1707009114