A k-mer based approach for virus classification identifies coronavirus infections and viral associations in human and plant microbiomes

Benjamin Garcia (benjamin.garcia@ucdenver.edu)
Oak Ridge National Laboratory https://orcid.org/0000-0001-5524-6946
Ramanuja Simha
Oak Ridge National Laboratory
Michael Garvin
Oak Ridge National Laboratory
Anna Furches
Oak Ridge National Laboratory
Piet Jones
Oak Ridge National Laboratory
Joao Gazolla
Oak Ridge National Laboratory
P. Doug Hyatt
Oak Ridge National Laboratory
Christopher W Schadt
Oak Ridge National Laboratory
Dale Pelletier
Oak Ridge National Laboratory
Daniel Jacobson
Oak Ridge National Laboratory

Methodology

Keywords: microbiome, metagenomic, COVID-19

Posted Date: October 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52940/v2

Abstract

Background Viruses are underrepresented taxa in the study and identification of microbiome constituents; however, they play an essential role in health, microbiome regulation, and transfer of genetic material. Only a few thousand viruses have been isolated, sequenced, and assigned a taxonomy, which further limits the ability to identify and quantify viruses in the microbiome. Additionally, the vast diversity of viruses represents a challenge for classification, not only in constructing a viral taxonomy, but also in identifying similarities between a virus' genotype and its phenotype. However, the diversity of viral sequences can be leveraged to classify their sequences in metagenomic and metatranscriptomic samples. Methods To identify and quantify viruses in transcriptomic and genomic samples, we developed a dynamic programming algorithm for creating a classification tree out of 715,672 metagenome viruses. To create the classification tree, we clustered proportional similarity scores generated from the k-mer profiles of each of the metagenome viruses. We then integrated the viral classification tree with the NCBI taxonomy for use with ParaKraken (a parallelized version of Kraken), a metagenomic/transcriptomic classifier. The resulting Kraken2 database of the metagenomic viruses can be found here: https://www.osti.gov/biblio/1615774 and is compatible with Kraken2. Results To illustrate the breadth of our utility for classifying viruses with ParaKraken, especially samples without virus-induced pathophysiology, we analyzed data from a plant metagenome study identifying the differences between two Populus genotypes in three different compartments and on a human metatranscriptome study identifying the differences between Autism Spectrum Disorder patients and controls in post mortem brain tissue. In the Populus study, we identified genotype and compartment-specific viral signatures, while in the Autism study we identified a significant increase in abundance of eight viral sequences in post mortem brains. We also show the potential accuracy for classifying viruses by utilizing both the JGI and NCBI viral databases to identify the uniqueness of viral sequences. Finally, we utilize the NCBI databases to identify pathogenic viruses in known COVID-19 and cassava brown streak virus infection samples to validate the potential usefulness of classifying viruses. Conclusion Viruses represent an essential component of the microbiome. The ability to classify viruses represents the compulsory first step in better understanding their role in the microbiome. Our viral classification method allows for a more complete identification of viral sequences than previous methods. This will improve identification of associations between viruses and their hosts as well as viruses and other microbiome members and can be used with any tool that utilizes a taxonomy for classification (such as Kraken).

Background

The number of phages on Earth is estimated to be as high as 4.80×10^{31} [1], implying the total number of viruses that might exist is much greater. Despite the high number of viruses, only a small number of viruses have been sequenced or characterized. Metagenomics and metatranscriptomics have led to the identification of large numbers of viruses; however, knowledge of their taxonomy is limited, making accurate identification in -omic samples challenging. While updates have been made to classify viruses by the International Committee on Taxonomy of Viruses (ICTV) [2], currently only 5,560 viral species have

been assigned a taxonomy. In contrast, the Joint Genome Institute's (JGI) Integrated Microbial Genomes & Microbiomes (IMG) [3] reports over 8,000 viral isolates and IMG/VR [4] lists more than 715,000 metagenomic viruses, the majority of which are devoid of taxonomic classification because they are sampled from a mixture of organisms and have not been isolated. Metagenomic viruses often lack phenotypic characteristics as well as host information, which creates challenges for understanding their basic biology [5] and requires different methodologies for classification [6]. Despite taxonomic and biological hurdles, identifying viruses in meta-omic experiments allows for novel insights into host-virus interactions, in addition to other interactions throughout the microbiome and phytobiome.

JGI's effort to assemble metagenomic viruses [7] has led to an unprecedented number of viral sequences that can be utilized to classify the microbial dark matter that can make up the bulk of metagenomic and metatranscriptomic samples. To the best of our knowledge, our k-mer based approach is the only extant method able to classify and quantify viruses at the scale provided by IMG-VR. One of the major challenges of classifying viruses, especially in the absence of phenotypic information, is their diversity [8] and the poor relationship between sequence and evolution [9]. While the inclusion of highly divergent sequences increases the difficulty of creating a detailed, fine-scale viral taxonomy, the presence of unique sequences can aid in the classification of viruses in -omic samples, i.e. we know they are likely different at the species level or greater. Methods such as natural vector representation [10][11], pairwise sequence comparisons [12], and pairwise evolutionary distances [13] have been developed to better identify phylogenetic similarity among viruses, but k-mer-based methods can provide the speed and scale that is necessary for highly efficient and accurate classification of millions to billions of sequencing reads against databases of taxonomic sequences [14].

In this paper, we create a methodology for generating a classification tree of 715,672 viruses from IMG-VR [4] for use in identifying viral sequences in metagenomic and metatranscriptomic studies (Figure 1). Given the infeasibility of comparing all viruses to each other, we first subset the viruses to identify which pairs have k-mer overlaps for calculating similarity scores. Subsetting resulted in the reduction of the comparison space by 99.98%, allowing for quantitative proportional similarity coefficients [15] to be calculated for each virus pair with a non-zero similarity. The algorithm Hip-MCL [16] was then used to cluster similar viruses for use in generating a hierarchical tree based on multiple inflation values (Supplemental File 1). The classification tree was integrated with NCBI's taxonomy, allowing for taxonomic classification of reads from metagenomic and metatranscriptomic samples. Finally, the pseudo-taxonomy was used to create Kraken [14] databases of all the metagenomic viruses for use in ParaKraken [17]. To illustrate the broad use of our viral classification method, we applied our ParaKraken databases to a plant dataset containing rhizosphere, endosphere, and soil samples of two different *Populus* genotypes and a dataset of post mortem brain samples from individuals with Autism Spectrum Disorder (ASD) and controls [18]. We also utilize NCBI's viral sequences as a positive control to identify how unique viral sequences are to the individual isolate and to identify known viruses under infection conditions, such as COVID-19 infection in a bronchoalveolar lavage (BAL) sample [19] and cassava brown streak virus in a cassava sample [20].

Results

The virus classification tree can be integrated into the NCBI taxonomy for classifying viral sequences from metagenomic and metatranscriptomic samples. To illustrate the power of having a classification tree of 715,652 viruses, we ran ParaKraken [17] on two different studies. The first study is a *Populus* metagenome that consists of two different genotypes (*P. deltoides* and *P. deltoides* x *P. trichocarpa*) in three different compartments (endosphere, rhizosphere, and soil). Each genotype-compartment pair had five replicates (with one *P. deltoides* endosphere sample removed for quality control reasons). Without the viral databases, the resulting metagenomic classification had 50% of reads with unknown taxonomic classification. The high amount of unknown microbial dark matter allows for a low-end estimation to be made of how many viral sequences may exist in microbiome samples. The second study is an ASD study comparing post mortem brain biopsies between ASD patients and controls. While the vast majority of reads in the ASD study are human, we captured a portion of the brain microbiome, including the viriome.

Populus **Metagenome**

To identify differences in the phytobiome between compartments and genotypes, we ran ParaKraken on endosphere, rhizosphere, and soil samples from *P. deltoides* and *P. deltoides x P. trichocarpa* (a hybrid). To mirror commonly used mapping methods in recent publications aimed at characterizing plant viriomes [21][22], the samples were initially analyzed using only the NCBI databases of all publicly available genomes from prokaryotes, archaes, eukaryotes and 8,000 viral taxa. The NCBI databases resulted in 50% of the 4.8 billion reads being assigned to taxa, with the rest representing unknown microbial dark matter (Fig. 2A). To illustrate the increased mapping ability of the k-mer based viral databases, we then analyzed the samples using our k-mer based viral databases in combination with the NCBI databases. The inclusion of the 715,672 JGI metagenome viruses resulted in an increase in mapping by 347 million reads. Metagenome viruses make up between 6%-20% (mean 15%) of the total mapped reads for a given sample, greatly increasing the coverage of the viriome (Supplemental Table 1). In addition to the increased read coverage, viral sequences also differ between compartments and genotypes leading to potential associations between viruses, the host, and other community members (Supplemental Table 2).

To better explore the differences between compartments and genotypes, we ran differential abundance analyses comparing both within and between genotypes, focusing on viruses. For the within genotype comparison, we compared endosphere vs rhizosphere, rhizosphere vs soil, and soil vs endosphere for each genotype (Fig. 3A). There were 65 viral sequences that were significantly differentially abundant across the comparisons. Rhizosphere (48 and 37 sequences) and soil (36 and 39 sequences) had much higher amounts of differentially abundant viral sequences compared to the endosphere (four and six sequences), likely due to the lower amounts of detected viral reads in the endosphere. Rhizosphere and soil profiles are similar to each other, with only 10 significant sequences unique to one compartment and genotype. Additionally, rhizosphere and soil samples within a genotype are more similar than rhizosphere and soil across genotypes. The two endosphere samples shared three out of the six unique significant endosphere viral sequences.

In addition to the within genotype comparisons, a within compartment comparison was performed as well comparing: *P. deltoides* soil to hybrid soil, *P. deltoides* rhizosphere to hybrid rhizosphere, and *P. deltoides* endosphere to hybrid endosphere (Fig. 3B). There were 48 significantly differential viral sequences across the comparisons. Similar to the within genotype comparison, the rhizosphere (28 and 15 sequences) and soil (27 and 16 sequences) had more significant viruses compared to the endosphere (three and two sequences). Additionally, the soil and rhizosphere samples for a given genotype have nearly identical significant viral sequences with one virus unique to the *P. deltoides* rhizosphere, three unique to the hybrid rhizosphere, and four unique to the hybrid soil. There were no significant viruses shared across genotypes for the rhizosphere and soil. The endosphere samples had no unique viruses associated only with the endosphere, and each endosphere sample shared at least one significant virus with the hybrid rhizosphere/soil and *P. deltoides* rhizosphere/soil. The significant differences in viral sequences associated with the *P. deltoides*, the hybrid, and the different compartments suggest host factors and differences in microbiome community composition may select for different viruses. Whether the lower number of viruses associated with the endosphere samples relative to the rhizosphere/soil is due to differential centrifugation, some host-related factor, or some database bias is unknown.

Autism Spectrum Disorder Metatranscriptome

To better understand the association between viral sequences and human health, we analyzed post mortem brain tissue metatranscriptomic samples [18] from ASD individuals and controls (Fig. 2B). We first aligned reads to the GRCh38 human reference genome, resulting in 67.5% (2.99 billion of the total 4.43 billion) of reads mapping to the reference. The unmapped reads were then processed with our ParaKraken pipeline. Unsurprisingly, 95% of the unmapped reads were assigned to eukaryotes, most likely due to ambiguous mappings and differences between the patient and the human reference genome (Supplemental Table 3). Despite the high percentage of human sequences in the unmapped reads, metagenomic viral reads were readily identified in all samples, ranging from 5 k to 125 k reads (avg 0.06% of reads; whereas, bacteria make up 0.57% of reads). To assess the uniqueness of the JGI viruses, we quantified the percent of reads that mapped at the metagenome virus level. Only 8.9% of reads mapped at a level higher than the individual virus, suggesting that the JGI viruses are highly unique and unambiguous.

To further understand if there is an association between ASD and viral sequences in the brain biopsies, differential abundance was performed comparing ASD to controls. Eight metagenome viral sequences were significantly more abundant in ASD cases relative to controls at a > 2x fold change, compared to zero significantly more abundant in the controls at that fold change. When comparing all of the significant viruses (p-value < 0.05 and f-value > 0.9; irrespective of fold change), the average fold change of the nine viruses significantly higher in ASD was 2.23 (1.99–2.62) compared to 1.09 (1.00-1.16) for the five viruses significantly higher in controls, suggesting that brain tissue from those with ASD may have relatively higher abundances of viral sequences (at least for the viruses contained in our databases). Increased numbers of polyomaviruses such as those identified here have previously been reported in brain tissue of individuals with ASD [23], and the number of viruses within an individual has been shown

to be correlated with decreased neuropsychological development [24], supporting the idea that there is an association between ASD and the viruses present in brain tissue.

Assessment of Virus Uniqueness and Abundance

To assess the uniqueness of viral sequences, we downloaded an updated version of NCBI's viral taxonomy in Feb. 2020 [25]. We first computationally generated reads from all of NCBI's viral sequences with a length of 200 bp and a sliding window of one, resulting in 1.06 billion reads. We then ran these reads through ParaKraken on both the NCBI and JGI databases. A slight majority of the reads (53.4%) mapped only to the individual viral isolate in which the read was generated. A small percentage of the reads (8.3%) mapped to the root, which means that the viral reads either had homology to either another superkingdom or to the metagenomic viruses. The uniqueness of the reads within the NCBI's database suggests that much of the viral diversity is undiscovered, there are few viruses in which there are multiple similar isolates sequenced, and there are few overlaps between NCBI's viruses and the viruses from JGI. The JGI viruses in the Autism metagenome had much lower than expected ambiguity based on NCBI's results. The ratio of reads that mapped to the individual isolate compared to non-root viral reads was 1.4 (53.4 / 38.3) for the viruses in NCBI compared to the 10.2 (91.1 / 8.9) for the JGI metagenome virus in the Autism dataset.

To further understand the uniqueness of NCBI's viruses, we used two different databases to classify viruses in a COVID-19 BAL sample [19] (Supplemental Table 4). The first database consists of the original NCBI and JGI parakraken databases, while the second database includes the Feb. 2020 version of NCBI's viruses (which includes the SARS-CoV-2 genome that causes COVID-19). Without the isolate of interest, 26878 (0.2% of total reads) reads mapped to different *Coronavirinae*, with the top hit of 22238 being SARS coronavirus (the old coronavirus taxonomy). With the addition of the SARS-CoV-2 isolate, 62480 reads (0.5% of total reads) now mapped to *Coronavirinae* with the majority 62461 mapping to the specific virus taxa of interest: severe acute respiratory syndrome coronavirus 2. SARS-CoV-2 has an 89.1% sequence homology with another SARS-like coronavirus [19], which is partly why we were able to identify coronavirus reads without the exact isolate of interest. However, despite having a highly similar virus already sequenced, the inclusion of the exact isolate of interest increased the mapping by 2.3x (reaffirming the prior results that the majority of viral sequences are unique given the sparsity of the number of viruses that have been sequenced).

In addition to the COVID-19 sample, we also analyzed a cassava sample from Mozambique that was confirmed to be infected with a cassava brown streak virus [20] (Supplemental Table 5). We ran ParaKraken with the original NCBI database and the JGI databases on the sample and confirmed the presence of the virus of interest. ParaKraken identified 1942 reads (0.1% of total reads) associated with the cassava brown streak virus. As expected, neither the coronavirus nor the brown streak virus was identified by ParaKraken in the Autism brain samples or the plant metagenome samples. Both the SARS-CoV-2 and cassava brown streak viruses demonstrate that ParaKraken can identify viruses of interest during an active infection, and that viruses contain highly unique sequences, partially due to the lack of viruses sequenced and viruses with taxonomy.

Discussion

Identifying constituents that comprise the microbiome and their relationship to the host is crucial to understanding human health, plant health, and how microorganisms impact phenotypes in general. Current methodologies for microbiome analyses are largely focused on bacteria using 16S sequencing, fungi with ITS sequencing, and other organisms using taxonomy through metagenomic sequencing; however, very little work has been done to quantify and understand viral reads in metagenome and metatranscriptome samples outside of what can be achieved by the few viruses that have been isolated and assigned to a taxonomy.

Addressing current limitations in virus identification is critical because viruses play an important but understudied role in many biological systems. For example, bacteriophages can modulate the metabolome of gut microbiomes in mice, including influencing the bacteria that the phage does not directly impact [26]. Additionally, end-stage viral infection in chimpanzees can cause a destabilization of the bacteria in the gut microbiome, likely through alteration of the host immune system [27]. The microbiome can also play a protective role, helping to decrease the risk of viral infections [28]. Much of a virus' role in microbiome samples is unknown due to the lack of methodology for their detection and the fact that the majority of information gleaned from microbiome samples is from bacteria.

Additionally, prior to this paper, there has been a lack of methodology to quantify viral taxa, study host-virus interactions, or interactions between viruses and other microbiome constituents at large scales. Directly measuring known viruses, viral antigens, and viral antibodies in human samples has led to the identification of associations between polyomaviruses and ASD [23], other viruses with neurodevelopment in general [24], different herpes viruses with multiple sclerosis [29] and peripheral neuropathies [30], HIV and peripheral arterial disease [31], and hepatitis C and kidney disease [32], etc. It is highly likely that there are many human-related and microbiome-related viruses that have implications for human health that await discovery. In plants, pathogenic viruses have been shown to increase the severity of drought through an increase in both infection and drought severity [33]; however, different infectious viruses have been shown to decrease drought severity by decreasing water loss, ultimately leading to improved tolerance of both infection and drought [33, 34]. Furthermore, the effect of combined viral stress and abiotic stressors (such as drought, heat, salinity, etc.) are affected by the overlap in epigenetic responses to individual stressors, which can produce positive or negative effects for both stressors depending on the often plant-specific individual responses to each stress [35]. The wide variety of direct and indirect effects of viruses necessitates both a better understanding of viruses and a better way of identifying viruses on large scales.

IMG/VR [4, 28] offers a suitable starting point for methodology development and hypothesis generation in identifying unknown viruses and their associations by providing the most extensive known collection of metagenomic viruses. While their work identifies viral assemblies in a single sample, we have utilized their large collection of assemblies to quantify viral sequences in diverse metagenomic and metatranscriptomic samples. To classify viral sequences, we have developed a dynamic programming

algorithm that allows one to create a classification tree from metagenomic viruses that can be integrated with NCBI's taxonomy. While it is impractical to compare all viruses to each other, the method initially identifies which viruses have a non-zero similarity score, reducing the number of similarity calculations by 99.98%. The reduction makes the calculation of similarities and construction of the classification tree feasible, providing the first step in identifying viral sequences in metagenomic and metatranscriptomic samples.

The classification tree's utility is demonstrated through identification of viral sequences in *Populus* genotypes and compartments and ASD and control brain biopsies. While the individual virus counts identified in the two datasets are orders of magnitude lower than cassava brown streak virus and COVID-19 infection samples, neither the *Populus* nor ASD samples presented with any known virus-induced pathophysiology. However, the *Populus* metagenome had a total viral load ~ 15% of reads on average, which is higher than the 0.3% of reads seen in the cassava sample with active infection. The higher total viral counts in a non-infected sample could be explained by the fact that viruses play many roles in the microbiome that are not related to active infection; as such, individual counts may not need to be as high as an infectious virus to have an impact on the microbiome and on the host.

Conclusion

Viruses are a vastly understudied component of microbiomes. The method we present here for creating a classification tree from metagenomic viruses can be utilized with any taxonomy-based classification tool to better identify viruses and their impacts in the microbiome. Although the 715,672 metagenome viruses that JGI has identified potentially make up only a small fraction of viruses that exist, we show that it is possible to identify viral sequences in metagenomic *Populus* genotype and compartment samples and metatranscriptomic ASD samples. More specifically, we identified eight significant differential viral sequences that are significantly higher and with a FC > 2.0 in ASD patients than in controls. We also show that our method can accurately identify viruses by utilizing NCBI's viral genomes to identify known viruses in COVID-19 and cassava brown streak virus infection samples. Through the use of NCBI's viral databases we also show that viral sequences are highly specific to the individual viral isolate, and that the JGI metagenome viruses have a higher uniqueness than the NCBI viruses. In addition to classification and quantification, further downstream analyses on viral reads, such as assembly, homology, functional annotation, etc. can be performed to predict features of the potential virus or viral sequence. Ultimately, a better understanding of the effects that viruses have on both the microbiome and the host will lead to a better comprehension of human health and plant biology.

Methods

Virus and Taxonomy Downloads

Metagenomic viruses were downloaded from JGI-IMG/VR [4] (N = 715,672 in January 2018). NCBI [25] whole genomes sequences containing 116 k (~ 67 k unique species/strains) eukaryotes, prokaryotes,

archaea, and viruses were downloaded in November 2017. The resulting classification tree from the metagenome viruses was then integrated with NCBI's taxonomy, which can be used with Kraken [14], ParaKraken [17], or any other classification method that utilizes NCBI's taxonomic tree.

Calculating Similarity

K-mers of size 31, using a sliding window of size one, were utilized to generate the classification tree as it allows for direct incorporation with our NCBI databases in Kraken or ParaKraken; however, the methodology presented here is agnostic to the k-mer size. Due to the inability to both store and quickly access all unique k-mers for all viruses (the number of unique k-mers was greater than the number of keys available in a hashtable), the viruses were first broken into 20 subsets. Each subset was compared to all other subsets, resulting in 71,566 viruses in each comparison set. First, all k-mers were generated for all viruses in each subset pair, with only the unique k-mers stored in a hashmap, along with any/all associated viruses containing that k-mer. Two k-mers were treated as the same if there was an exact match for the 31-mer, utilizing both the forward and reverse complement for the determination of an exact match.

The forward and the reverse complement of each sequence was used because viral genomes can be a mixture of DNA/RNA and single/double-stranded viruses, the strandedness of some viruses are different at different life cycle stages, and some viruses have a genome that is partially single and partially double stranded. While the metagenomic viruses are composed of DNA viruses, the methodology was developed to be agnostic to the type of virus. K-mers with ambiguous bases were not stored. K-mers with more than one virus associated with the sequence are indicative of overlap between two or more viruses. Viruses with at least one overlapping k-mer with another virus were stored in pairs to calculate similarity scores. Any virus pair without any k-mer overlaps were assigned a similarity score of zero. The initial subset decreased the number of similarity scores to be calculated from ~ 256 billion to only ~ 43 million (a decrease of over 99.98%), making the computation much more feasible. The decrease was achieved due to the sparsity of virus pairs that had any overlapping k-mers, resulting in the majority of virus pairs having a similarity score of zero.

Quantitative proportional similarity coefficients were calculated between the sets of k-mers for each virus pair with overlapping k-mers. Due to memory limitations of 500G per node and speed limitations of calculating all similarities in a linear fashion, the ~ 43 million overlapping virus pairs were broken into 400 subsets. K-mer profiles from each virus in each subset were generated beforehand to eliminate the need for regenerating profiles with every comparison (as a given virus can appear in multiple subsets). However, to decrease the memory overhead (at the cost of a longer run), the k-mer profiles can be generated for each virus and for each comparison. Each of the subsets was then run on the Summit supercomputer in parallel. K-mer overlap was calculated in the same manner previously with a k-mer and its reverse complement being treated as the same k-mer. The result is a matrix of quantitative proportional similarity coefficients for all virus pairs.

Virus Classification Tree

Groups of similar viruses were identified by running HipMCL [16], a parallel Markov clustering algorithm, on the triples of virus k-mer similarity scores. Inflation values of 1.4, 2.0, 3.0, 4.0, and 6.0 were utilized to identify a range of cluster sizes. The resulting clusters were integrated into a hierarchy by MCLCM [36]. Clusters of more than three viruses (excluding the 32,084 viruses that had no similarity to any other virus) were reclustered using neighbor joining to increase granularity, resulting in a hierarchy of viruses consisting of 22 levels (one of the cluster trees shown in Supplemental Fig. 1). Pseudo-taxonomic IDs were then assigned to each level and virus in order to integrate with NCBI's taxonomy. The integrated taxonomy was then used to create databases for use in ParaKraken. The Kraken2 database of the JGI metagenome viruses can be found here: https://www.osti.gov/biblio/1615774 [37].

Autism and Populus **Datasets**

Bulk RNA-Seq from 22 ASD and 19 control post mortem brain samples were obtained from Velmeshev et al [18]. FASTQ files were trimmed using Atropos [38] and mapped to the GRCh38 human reference genome using STAR [39]. *P. deltoides* and *P. deltoides* x *P. trichocarpra* endosphere, rhizosphere, and soil samples (five samples of each compartment and genotype combination; one *P. deltoides* endosphere was removed for quality reasons) were obtained from JGI (https://gold.jgi.doe.gov/biosamples?Study.GOLD+Study+ID=Gs0103573). Endosphere samples underwent differential centrifugation to decrease the concentration of plant host [40], and then all samples underwent paired-end sequencing. Reads were trimmed using skewer [41] and filtered against *P. deltoides* and *P. trichocarpa* reference genomes (~ 0.2% of reads aligned to the host). Unmapped reads from both the *Populus* and ASD datasets were run through ParaKraken [17]. A median normalization was applied, taxa with less than 75% coverage across samples were removed, taxa making up less than 0.01% of the reads at a species level in the ASD data and 0.001% of the reads at a species level in the *Populus* data were also removed. Differential abundance was then assessed using fcros [42] with p-value < 0.05 and f-value > 0.9, and networks were visualized using Cytoscape [43].

Cassava, Coronavirus And Updated Viral Genomes Datasets

A cassava root RNA-Seq sample with a confirmed brown streak virus was obtained from Amisse et al. [20] and ran through ParaKraken [17] to classify taxa. A BALsample with a confirmed COVID-19 infection was obtained from Wu et al. [19]. Reads were trimmed using Atropos [38] and then run through ParaKraken for taxa classification. We then downloaded the latest RefSeq viral genomes from NCBI in Feb 2020 [25], as the updated version contained COVID-19. We then reran ParaKraken on the Autism sample to compare the read counts for coronavirus pre- and post-COVID-19 genome inclusion. To assess the uniqueness of viruses in the Feb 2020 version of RefSeq, we created reads from all 67,519 RefSeq viruses using a read length of 200 and a sliding window of one nucleotide. The resulting 1.06 billion reads were then run through ParaKraken. The taxonomic assignment was then compared to the taxonomy of the viruses in which the reads originated.

Declarations

Ethics Approval and Consent to Participate

Not Applicable

Consent for Publication

Not Applicable

Availability of Data and Materials

The Kraken 2 database created from the JGI metagenome data and used for identifying metagenomic viruses in the samples utilized in this study can be found: : https://www.osti.gov/biblio/1615774

Other data supporting the conclusions found in the manuscript can be found in the supplemental data as follows.

References

1. Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. Viruses as Winners in the Game of Life. Annu Rev Virol. 2016;3:197–214.
2. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, et al. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). Arch Virol. 2019;164:2417–29.
3. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012;40:D115–22.
4. Paez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2017;45:D457–65.
5. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15:161–8.
6. 10.1038/nprot.2017.063
   Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data [Internet]. Nature Protocols. 2017. p. 1673–82. Available from: http://dx.doi.org/10.1038/nprot.2017.063.
7. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536:425–30.
8. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15:161–8.

9. Simmonds P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol. 2015;96:1193–206.
10. Yu C, Hernandez T, Zheng H, Yau S-C, Huang H-H, He RL, et al. Real time classification of viruses in 12 dimensions. PLoS One. 2013;8:e64328.
11. Deng M, Yu C, Liang Q, He RL, Yau SS-T. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One. 2011;6:e17293.
12. Bao Y, Chetvernin V, Tatusova T. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. Arch Virol. 2014;159:3293–304.
13. Lauber C, Gorbalenya AE. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J Virol. 2012;86:3890–904.
14. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
15. Weighill DA, Jacobson D. Network Metamodeling: Effect of Correlation Metric Choice on Phylogenomic and Transcriptomic Network Topology. Adv Biochem Eng Biotechnol. 2017;160:143–83.
16. Azad A, Pavlopoulos GA, Ouzounis CA, Kyrpides NC, Buluç A. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Res. 2018;46:e33.
17. 10.1094/pbiomes-04-18-0021-r
Garcia BJ, Labbé JL, Jones P, Abraham PE, Hodge I, Climer S, et al. Phytobiome and Transcriptional Adaptation of Populus deltoides to Acute Progressive Drought and Cyclic Drought [Internet]. Phytobiomes Journal. 2018. p. 249–60. Available from: http://dx.doi.org/10.1094/pbiomes-04-18-0021-r.
18. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364:685–9.
19. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature Nature Publishing Group. 2020;579:265–9.
20. Amisse JJG, Ndunguru J, Tairo F, Ateka E, Boykin LM, Kehoe MA, et al. Analyses of seven new whole genome sequences of cassava brown streak viruses in Mozambique reveals two distinct clades: evidence for new species. Plant Pathol. 2019;68:1007–18.
21. Jo Y, Lian S, Chu H, Cho JK, Yoo S-H, Choi H, et al. Peach RNA viromes in six different peach cultivars. Sci Rep. 2018;8:1844.
22. 10.1128/JVI.01462-19
Ma Y, Marais A, Lefebvre M, Theil S, Svanella-Dumas L, Faure C, et al. Phytovirome Analysis of Wild Plant Populations: Comparison of Double-Stranded RNA and Virion-Associated Nucleic Acid Metagenomic Approaches. J Virol [Internet]. 2019; Available from: http://dx.doi.org/10.1128/JVI.01462-19.

23. Lintas C, Altieri L, Lombardi F, Sacco R, Persico AM. Association of autism with polyomavirus infection in postmortem brains. J Neurovirol. 2010;16:141–9.
24. Karachaliou M, Chatzi L, Roumeliotaki T, Kampouri M, Kyriklaki A, Koutra K, et al. Common infections with polyomaviruses and herpesviruses and neuropsychological development at 4 years of age, the Rhea birth cohort in Crete, Greece. J Child Psychol Psychiatry. 2016;57:1268–76.
25. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
26. Hsu BB, Gibson TE, Yeliseyev V, Liu Q, Lyon L, Bry L, et al. Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. Cell Host Microbe. 2019;25:803–14.e5.
27. 10.1002/ajp.22515
Barbian HJ, Li Y, Ramirez M, Klase Z, Lipende I, Mjungu D, et al. Destabilization of the gut microbiome marks the end-stage of simian immunodeficiency virus infection in wild chimpanzees. Am J Primatol [Internet]. 2018;80. Available from: http://dx.doi.org/10.1002/ajp.22515.
28. 10.1093/cid/ciz968
Tsang TK, Lee KH, Foxman B, Balmaseda A, Gresh L, Sanchez N, et al. Association between the respiratory microbiome and susceptibility to influenza virus infection. Clin Infect Dis [Internet]. 2019; Available from: http://dx.doi.org/10.1093/cid/ciz968.
29. 10.1101/737932
Engdahl E, Gustafsson R, Huang J, Biström M, Bomfim IL, Stridh P, et al. Increased serological response against human herpesvirus 6A is associated with risk for multiple sclerosis [Internet]. Available from: http://dx.doi.org/10.1101/737932.
30. 10.1177/1941874414535215
Brizzi KT, Lyons JL. Peripheral Nervous System Manifestations of Infectious Diseases [Internet]. The Neurohospitalist. 2014. p. 230–40. Available from: http://dx.doi.org/10.1177/1941874414535215.
31. Beckman JA, Duncan MS, Alcorn CW, So-Armah K, Butt AA, Goetz MB, et al. Association of Human Immunodeficiency Virus Infection and Risk of Peripheral Artery Disease. Circulation. 2018;138:255–65.
32. Fabrizi F, Donato FM, Messa P. Association Between Hepatitis C Virus and Chronic Kidney Disease: A Systematic Review and Meta-Analysis. Ann Hepatol. 2018;17:364–91.
33. Cui Z-H, Bi W-L, Hao X-Y, Xu Y, Li P-M, Walker MA, et al. Responses of In vitro-Grown Plantlets (Vitis vinifera) to Grapevine leafroll-Associated Virus-3 and PEG-Induced Drought Stress. Front Physiol. 2016;7:203.
34. Ramegowda V, Senthil-Kumar M. The interactive effects of simultaneous biotic and abiotic stresses on plants: mechanistic understanding from drought and pathogen combination. J Plant Physiol. 2015;176:47–54.
35. Pandey P, Ramegowda V, Senthil-Kumar M. Shared and unique responses of plants to multiple individual stresses and stress combinations: physiological and molecular mechanisms. Front Plant

Sci. 2015;6:723.

36. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.
37. Garcia BJ, Simha R, Garvin M, Furches A, Jones P, Hyatt PD, et al. Kraken2 Metagenomic Virus Database [Internet]. 2020. Available from: https://www.osti.gov/biblio/1615774.
38. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017;5:e3720.
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
40. Utturkar SM, Cude WN, Robeson MS Jr, Yang ZK, Klingeman DM, Land ML, et al. Enrichment of Root Endophytic Bacteria from Populus deltoides and Single-Cell-Genomics Analysis. Appl Environ Microbiol. 2016;82:5698–708.
41. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;15:182.
42. Dembélé D, Kastner P. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. BMC Bioinformatics. 2014;15:14.
43. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

Figures

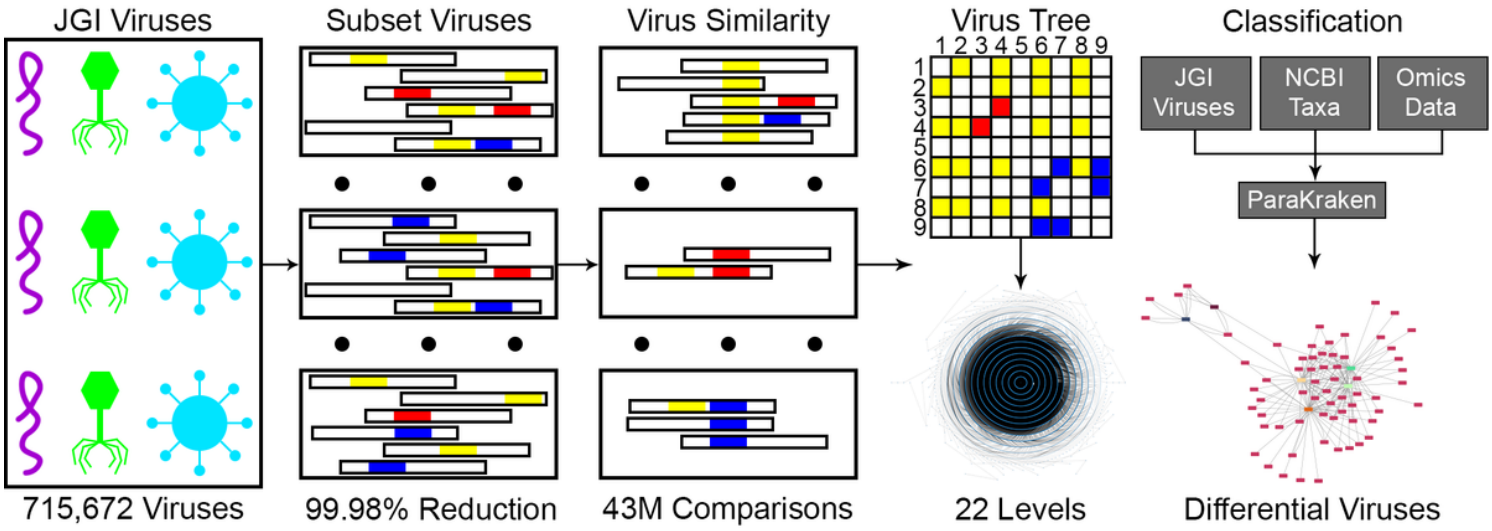

Figure 1

Creation and use of a viral classification tree. To first identify which of the 715,672 metagenome viruses have non-zero similarity scores with other viruses, we subset the viruses to identify k-mer overlaps. We identify ~43 million pairs with nonzero similarity scores, a reduction in the number of calculations by

99.98%. Clusters of viruses are then created by running HipMCL on quantitative proportional similarity coefficients with the following inflation values: 1.4, 2.0, 3.0, 4.0, and 6.0. MCLCM was then run on the different inflation clusters to generate a hierarchy, and then neighbor joining was run on clusters with more than 3 members to add more structure to the trees. The metagenome virus tree was then integrated with NCBI's taxonomy for use in classifying metatranscriptomic and metagenomic samples. We ran the NCBI whole genomes and JGI metagenome viruses with ParaKraken on a Populus and an ASD dataset, allowing us to classify taxa and identify differential abundance across conditions. The network is the result of a differential abundance of viruses (violet-red) within the Populus genotypes (green - hybrid, orange - P. deltoides). There is a similarity in virus abundances across the genotypes in the soil (lightest) and the rhizosphere (middle) with the endosphere (darkest) being dissimilar to the other compartments.

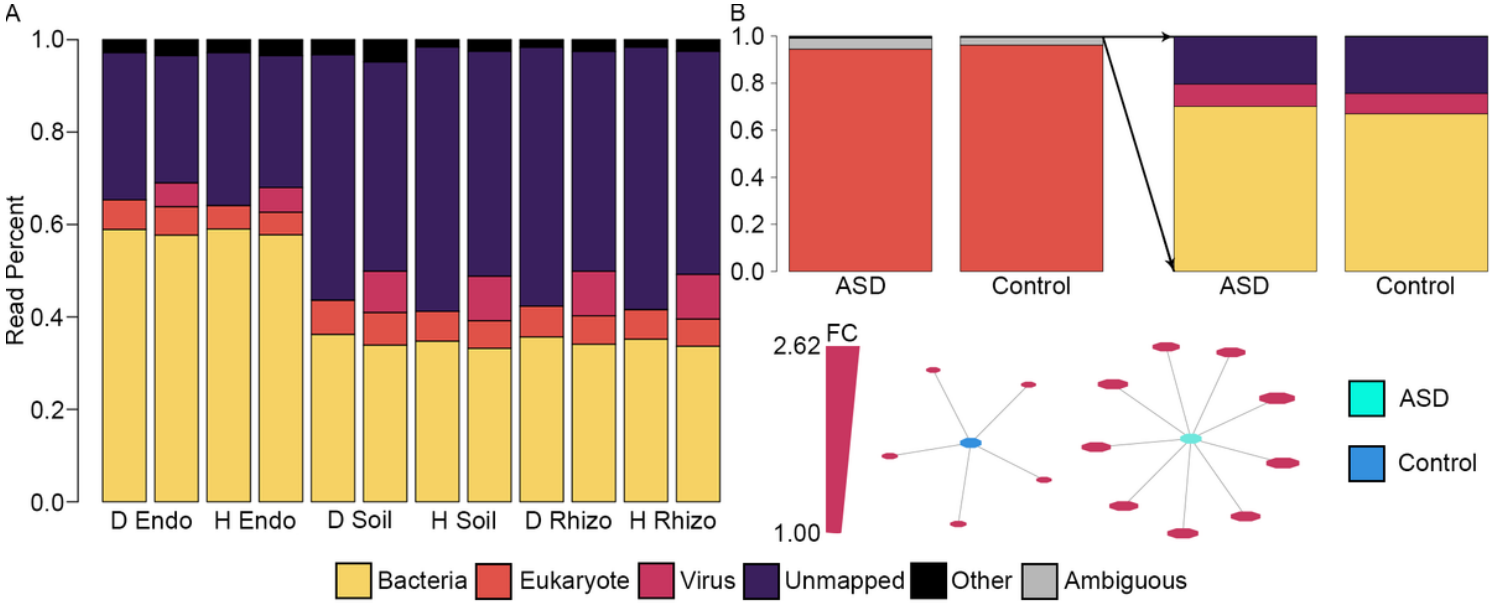

Figure 2

Classification of viruses in metagenomic and metatranscriptomic samples. A) Effect of the virus databases on the number of reads mapped in the Populus deltoides (D) and hybrid (H) data. The first bar in each group represents ParaKraken results before the viral databases, while the second bar shows the classification after the inclusion of the viral databases. Metagenome viruses averaged 15% of the mapped read with higher percent mapping in the rhizosphere (Rhizo) and soil (Soil) relative to the endosphere (Endo), suggesting that viruses can make up a substantial portion of the microbiome. B) Differences in viruses between ASD brains and control brains. Unsurprisingly, eukaryotes make up the vast majority of reads of human samples. However, we are still able to identify sequences associated with viruses, eight of which are significantly higher in ASD versus controls (>2 fold change). The graph shows all significant differential abundance viruses (irrespective of fold change), with sizes of the viruses representing fold change (smallest - 1.00 FC, largest 2.62 FC). While there were five viruses with p-value < 0.05 and f-value > 0.9 in controls, their average fold change was 1.09, compared to 2.23 for the nine viruses higher in ASD, suggesting ASD brains may have higher viral counts. Other - NCBI viruses, viroids, ambiguous sequences. FC - fold change.

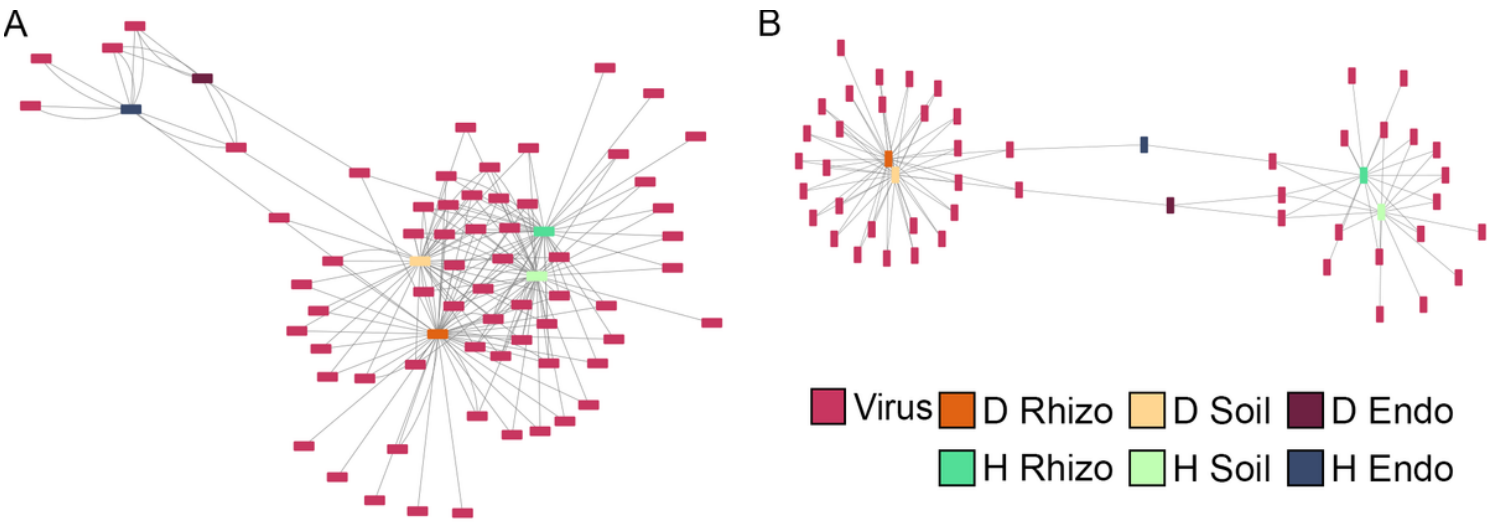

Figure 3

Differential abundance of viral sequences in Populus genotypes and compartments. In the A) within genotype comparison, the rhizosphere and soil samples have similar significant viral sequences across genotypes; however, the significant soil and rhizosphere viral sequences are more similar within genotypes than across genotypes. The two endosphere samples have few significant viral sequences, and they have more in common with each other than they do to the other compartments. In the B) between genotype comparison, the soil and rhizosphere samples for a given genotype have similar significant differentially abundant viruses. Additionally, the endosphere samples have much fewer significant differential sequences compared to the rhizosphere and soil, likely due to the overall lower abundance of viral sequences. Both graphs suggest there is a host or microbiome mediated selection of viral sequences that has some genotype and compartment specificity. D - P. deltoides, H - hybrid, Endo - endosphere, Rhizo - rhizosphere.

Supplementary Files

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

- SupplementalTable5.xlsx
- SupplementalTable4.xlsx
- SupplementalTable3.xlsx
- SupplementalTable2.xlsx
- SupplementalTable1.xlsx
- metavirussupFig1.tif
- virustreeparentcomplete.txt