Title : Viral and eukaryotic communities of urban ecosystems across US metropolitan areas

Authors: Serghei Mangul[#1], Nathan LaPierre[1], Igor Mandric[2], Lana Martin[1], Nicholas Wu[1], Eleazar Eskin[1], David Koslicki[3]

Affiliations:
[1] University of California, Los Angeles
[2] Georgia State University
[3] Oregon State University

# Correspondence to: Serghei Mangul smangul@ucla.edu

Microorganisms are ubiquitous in almost every ecosystem on earth, including soil, seawater, and the human body. Microscopic single- and multi-celled eukaryotes play many vital roles in their host environments (Handelsman 04, Wooley 10). Identifying the microbes present in a sediment, seawater, surface swab, or human tissue sample is critical to understanding what functions are carried out by these organisms. A more comprehensive characterization of microbial ecology will reveal how disturbances in microbial communities can lead to infection and disease in plants, humans, and other animals.

Microorganisms are traditionally studied using culture-based techniques, in which the microbial organisms are isolated from samples and individually studied in the laboratory setting. Culture-based techniques are incapable of capturing the complex relations that take place between the hundreds to thousands of different microbial communities in their natural habitats (Handelsman 04, Wooley 10). Recently, high-throughput sequencing has revolutionized microbiome research by enabling the study of thousands of microbial genomes directly in their host environments. This approach, which forms the field of metagenomics, avoids the biases incurred with traditional culture-dependent analysis. The metagenomics approach also allows the comparison of microbial communities' composition in their natural habitats across different human tissues and environmental settings (Handelsman 04, Wooley 10). Specifically, metagenomic profiling is proven useful for analyzing microbes such as eukaryotic and viral pathogens, which were previously impossible to study in an unbiased way with target 16S ribosomal RNA gene sequencing (Venter 04, Hugenholtz 08, Rosario 11).

Several existing methods for metagenomic profiling propose using 'marker genes' to identify the species present in a sample. Studies show this method can be efficient and accurate when estimating the presence and relative abundances of bacteria and archaea in a sample (Liu 11, Segata 12, Troung 15). However, approaches based on marker genes have some limitations when identifying viral and eukaryotic genomes. For example, one approach uses genes that are considered 'universal' to detect bacterial taxa (Liu 11). However, this approach does not accurately identify viruses, which are comprised mostly of novel sequences and do not share any single common gene (Willner 13, Wooley 10, Edwards 05). Another approach utilizes genes that uniquely identify a given phylogenetic clade (Segata 12, Troung 15). This approach is limited by the relatively small number of reads that map to the study's specified regions of the genome (Wood 14) and leads to poor sensitivity (Sczyrba 17). This approach also has poor read utilization when detecting eukaryotic genomes, which are usually long and comprised mostly of noncoding regions (Gilbert 11, Hugenholtz 08, Cowan 05). Recent approaches based on k-mers may overcome these issues and dramatically improve run time (Wood 14), but users have encountered issues with memory usage and sensitivity (Corvelo 17, Sczyrba 17).

We recently developed computational method, miCoP, capable of profiling many of the microorganisms in a sample—including bacteria, archaea, fungi, plasmids, and other single- and multi-cell eukaryotes. In contrast to traditional methods, which use only a small fraction of reads sequenced in a sample and maps the subsampled reads on to the marker genes. This method attempts to use all the reads by mapping them onto the most complete compendium of reference genomes. These advantages allowed us to use miCoP to study eukaryotic communities present in public spaces across urban ecosystems. To do this, we first combined reference databases for all of these organisms and mapped the shotgun sequence reads to these databases. Computational speed has traditionally been the primary limitation for mapping metagenomic sequences, but recently-developed ultrafast alignment methods and improvements in computing hardware have made miCoP computationally feasible. We use the protein alignment tool Diamond, which is 20,000 times faster than BLASTX with similar sensitivity. Since Diamond aligns genomic reads to a protein sequence database, we perform a six-frame translation on our genomic databases. After alignment is performed with Diamond, we use a quality filtering method to eliminate low-quality matches. Finally, we estimate the relative abundance of each species in the sample. Results of running miCoP on the simulated datasets show that our method outperforms current state-of-the-art methods by identifying more species in a sample, identifying low-abundance species, and utilizing more reads overall.

As viral genomes are typically much smaller than their bacterial, archaeal, and fungal counterparts, we utilized a recent computational method we developed, called Containment Min Hash (CMH) based on the min hash estimate of the Jaccard index (Ondov 2016), but which has significantly improved sensitivity and specificity (all computed Jaccard indices 95% confident to have a relative error of less than 0.10). After utilizing CMH to detect the presence of viral organisms (obtained from NCBI), we utilized the SNAP DNA-DNA aligner (Zaharia 2011) to align the samples to the detected organisms (all reads with MAPQ greater than 20).

To analyze microbiome communities in urban areas, we used the metagenomics data provided in the MetaSUB Inter-City Challenge (CAMDA 2017). The dataset was comprised of metagenomics samples collected from subway and light rail stations across three US metropolitan areas: New York (n=1572), Boston (n=141), and Sacramento (n=18). We randomly selected 65 samples from New York subway stations and considered all samples from Sacramento and Boston stations. In total, 224 samples were available for the analysis of microbiome composition and diversity across US metropolitan areas. In our analysis, we detected a significant number of reads originating from human endogenous retroviruses and reads originating from phiX174 enterobacteria phage (PhiX174 is routinely used as a part of the sequencing protocol), and so filtered out these reads.

First, we study virome composition within each subway and light rail station, and we compare composition across stations. Using CMH, and a stringent threshold (Jaccard index >= 10e-5, viral genome coverage >=50%), we were able to detect 27 viruses present in at least one subway station. Relaxing this threshold (Jaccard index >= 10e-7), we detected 334 viruses present in a least one subway station. Among the viruses, we detect skin-associated viruses, including *Molluscum contagiosum* (water warts). Among other viruses, we detect numerous phages, including *Bacillus* prophage, *Erwinia* phage, *Shigella* phage, and others. Figure 1 contains a depiction of a selection of viruses for which one of the samples had a significant number of high-quality reads aligning to the virus.

In general, we observe that the urban virome (a consortium composed of vast arrays of viruses) is station- and area-specific. This clustering effect can be partially attributed to the smaller fraction of reads originating from a virome compared to the total number of reads originating from metagenomic samples. For example, the sample from the Boston subway system has a low coverage median of 48,000 reads per samples, resulting in a cluster of samples in which no non-contaminant viruses were found (see Figure 1 a). In comparison, both the New York and Sacramento metagenomics samples have millions of reads available for study and so many more viruses were detected in these samples.

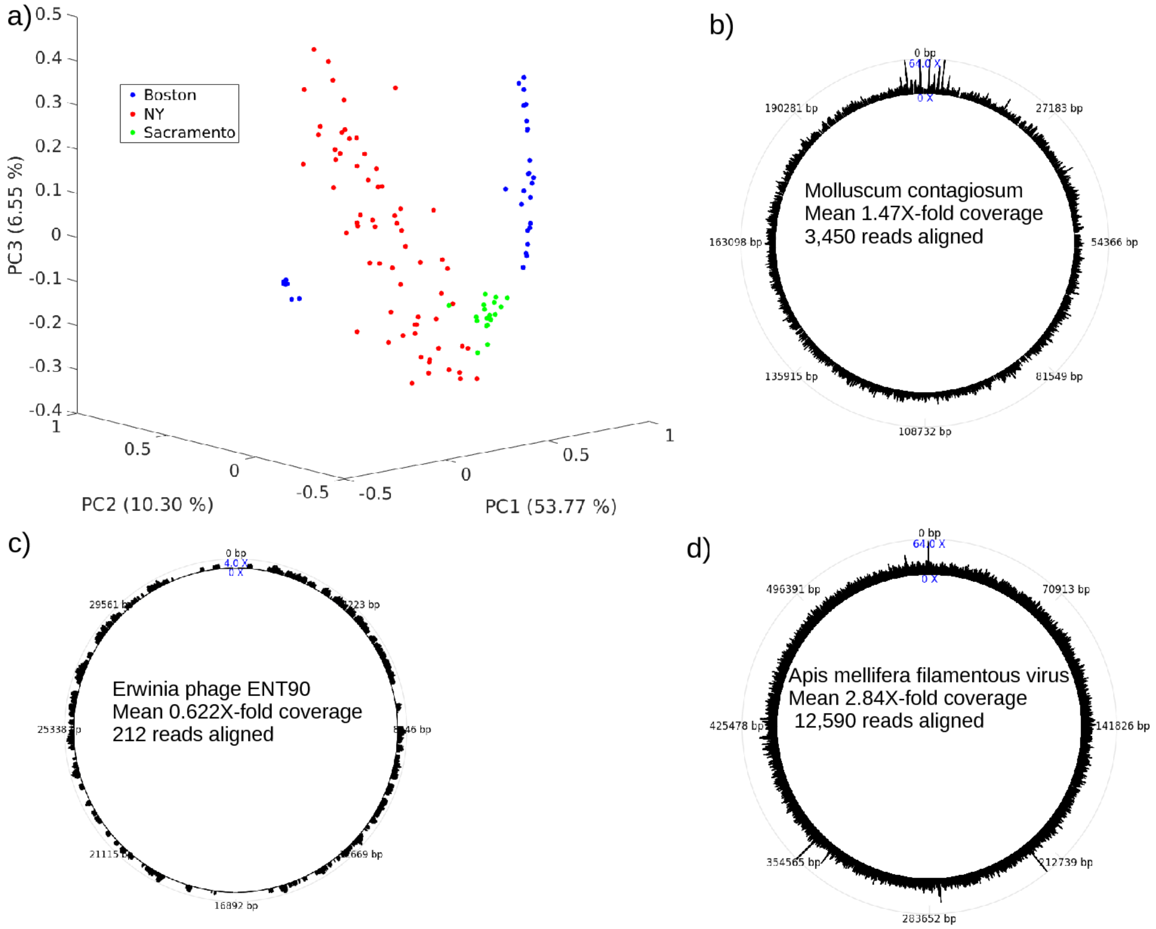

Figure 1. Composition of urban virome across US metropolitan areas. a) Cosine similarity PCoA plot of Jaccard indicies. b)-d) Alignment coverage plots of the most frequently covered viral organism. Outer ring is viral reference genome location, heights indicate the number of reads aligned using a root scale using 10-bp non-overlapping windows. b) Skin-associated virus present in one station of Boston subway system, c) New York subway system, and d) Sacramento light rail system.

Next, we used miCoP and applied similar criteria and filtered fungus organisms with less than 100 reads support. With all three urban areas combined, this approach detected 2788 distinct fungus species present in at least one subway or light rail station. Some of those species were previously reported to be part of the urban ecosystem in the Boston metropolitan area (Hsu, Tiffany, et al. 2016). For example, the *Malassezia* fungus is widely present within the metropolitan area ecosystem (and present across the North American continent). This fungi, which grows on the sebaceous areas of human skin, was supported on average with 2 million reads per sample across 92 subway stations (2% relative abundance).

In each of the 224 samples, the most widely-spread fungus present is *Pythium aphanidermatum*. This soil borne plant pathogen is from the taxonomic class known as water molds (Oomycetes). *Pythium aphanidermatum* was supported by 13459 reads on average per sample. Another common fungus identified in the samples is from *Aspergillus* genus. These black mold colonies are commonly reported in indoor urban environments (Samson, Robert A., et al. 2002).

In addition, we observe a widespread presence of several common plant and soil fungi across all three subway stations. Some fungi were present in surface swabs at many of the subway stations. We identified *Albugo laibachii*, a plant fungus that causes the host plant to become more susceptible to other parasites, in samples from 179 subway stations across three metropolitan areas. Other fungi were city- and station- specific. For example, a *Coccidioides immitis* (valley fever), a pathogenic fungus known to reside in the soil of southwestern regions of the US was detected at the surface of one of Sacramento light rail stations. We observe 60% of detected fungi to be present in more than 15 samples, with many of those spread across the three metropolitan areas.

In conclusion, we used CMH and miCoP to successfully obtain the composition of urban virome and eukaryome from samples obtained from multiple subways stations across three US metropolitan areas. We observe that many viruses are station-specific, suggesting that composition of virome is shared by the location and district human populations relevant to the stations. On another hand, we found that the eukaryome was composed of several species that are widely spread across the vast majorities of the station. Given the wide geographical distribution of those species (e.g., *Aspergillus* and *Pythium aphanidermatum*), they can be characterized as typical North American urban eukaryome.