Elucidating the metagenomic diversity of and across New York City subway stations

Authors: Cecelia Pierce Haider[1], Alexa Rosenberg[1]
Mentors: Dr. Jonathan Foox[2], Lauren Mak[2], Krista Ryon[2], Namita Damle[2], Dr. Christopher Mason[2]
Cold Spring Harbor Laboratory, DNA Learning Center[1]; Weill Cornell Medical Center[2]

Abstract

The urban microbiome of the New York City subway interacts with over a billion riders a year and is incredibly diverse. In our study, we aimed to compare the DNA collected at two terminal stations in more sparse regions of NYC with DNA collected at dense transfer hubs. We performed two swabbing runs at each station, in the morning and afternoon, and at three different locations in the station, in order to observe the effect of time and location on the DNA collected and sequenced at each station. Our results demonstrated that species diversity was greater at both transfer stations than both terminal stations, and the swabbing location that yielded the largest number of classified DNA reads was the hand railing. These results are significant because characterizing the urban microbiome of the NYC subway system can help with future public health endeavors and the creation of smarter cities.

Introduction

Metagenomics is a discipline in which the genomic content of all of the microorganisms of a particular niche are characterized (Handelsman et al. 1998). An analysis of the metagenome of an environment can be done both on natural environments such as water from a lake or sediment from a beach, as well as built environments. Built environments are defined as generally as all structures built by humans, such as a house, a library, or, in our case, a transit system (Gilbert et al. 2018). While built environments and other ecosystems both contain microbes, those in built environments are mostly spread by humans. Scientists are interested in learning more about these microorganisms because they play a major role in the cause and prevention of diseases in humans (Gilbert et al. 2018). Research on the metagenomics of urban environments is important because it creates a molecular/microbial map which can be used to determine how these built environments affect human health (Danko et al. 2021).

Millions of people ride the New York city subway every day, transmitting and admixing countless microbes among them, and the number of riders is increasing as New York recovers from the pandemic (https://new.mta.info/coronavirus/ridership). Profiling the metagenome of this densely populated built environment is essential foundational work for public health and for monitoring future outbreaks of pathogens and microbial transmissions as they spread around the world (Zhu et al., 2017). For this reason, we propose to conduct a metagenomic study of the New York City subway and to contribute to the efforts started by Danko et al. (2021) to characterize this built environment.

Metagenomic studies have been performed on hospitals (Brooks et al., 2017; Lax et al., 2017), soil (Hoch et al., 2019; Joyner et al., 2019), sewage (Fresia et al., 2019; Maritz et al., 2019), and transit systems (Afshinnekoo et al., 2015; Hsu et al., 2016; Kang et al., 2018; Leung et al., 2014; MetaSUB International Consortium et al., 2016). Many of these other studies performed on urban built environments have been done as part of the International Metagenomics and Metadesign of Subways and Urban Biomes Consortium (also known as MetaSUB). MetaSUB is a global network of scientists that share ongoing data about public transportation in urban environments in order to better understand the metagenomics of cities and their transit systems as well as their impact on humans. The pilot MetaSUB study (Danko et al. 2021) described geographical variations and drew connections between different cities and their metagenomes. These findings could reflect epidemiology and perhaps even have forensic applications for source-tracking. However, due to the scope of the project, the initial study did not differentiate between microbial profiles in different station types, which can accommodate very different numbers and mixtures of people.

In our study, we investigated the relationship between the metagenomes of terminal stations and transfer stations in the New York City subway system in order to compare the microbial profiles of different stations. We hypothesized that if we compared the metagenomes of terminal and transfer stations, then the transfer stations would show greater species diversity while the terminals would be more genetically isolated.

Acknowledgements

Thank you to our mentors Dr. Foox, Dr. Mason, Lauren Mak, Namita Damle, Deena Najjar, and Krista Ryon.

Thank you to the Pinkerton Foundation and Science Sandbox: an Initiative of the Simons Foundation.

Thank you to Dr. Allison Mayle, Arden Feil, and the rest of the Urban Barcode Research Program.

Methods

Samples were collected from four stations; the 96th street Q station, Times Square, Grand Central, and Flushing Main Street once in the morning and once in the evening using standard Metasub sample collection protocol. Collectors of samples wore gloves in order to avoid contamination. At each station, a ticket machine keypad, a door handle, and a hand railing were swabbed. These locations were chosen because they were surfaces that could be found in every station and they were frequently touched by passengers. Swabbing consisted of wetting the swab in a tube of 400 µL of DNA shield, vigorously rubbing an isohelix swab on an area of around 1 square foot for 3 minutes, and then breaking the head of the swab off in the DNA shield tube. For negative controls, the same procedure was followed, but instead of swabbing a surface, the swab was waved in the air for 2 minutes. Date, time, location, tube number, and the surface were recorded. Samples were refrigerated until they were taken to the laboratory for analysis. Samples were collected at 4 different locations in 4 stations at 2 different times (Ryon, 2022).

A ZymoBIOMICS MagBead DNA Extraction kit was used to extract DNA from samples using ZymoBIOMICS protocol. The NEBNext Ultra II DNA Library Prep Kit for Illumina was used to prepare libraries using a PCR amplification and then a cleanup of the PCR reaction following the NEB protocol. Extracted samples were then pooled into equimolar mixtures and assessed for quality. High quality libraries were then loaded onto an Illumina iSeq for sequencing in the Mason Lab at Weill Cornell Medicine.

First, human reads were removed from each sample using a custom script for removing human reads. Next, adapter sequences were removed using a program called AdapterRemoval. Quality checks were generated using FastQC and samples of unsatisfactory quality were removed. Kraken, a k-mer based classification tool, was used to match reads to known species in the NCBI Taxonomy database and generate summaries. Finally, results were visualized using custom scripts.

Results

Before running quality checks on our data, we used the Weill Cornell Cluster to remove human reads and adapter sequences from our sequencing data so that we could look at only the genetic material from microbes, fungi, protists, plants, and even other animals. We then used FastQC to generate a quick quality check on our sequencing data. We found that every sample except for three reads had satisfactory per base sequence quality, per tile sequence quality, and per sequence quality scores. The per base sequencing quality was passable for most samples, but some had lower certainty scores. We did not expect this to be a big problem in our sequencing because we used an Illumina sequencer, which is able to generate data on much smaller volumes of material.

Many samples had problems with per base sequence content and per base GC content. In fact, not a single sample had a satisfactory score in terms of per base GC content. Figure 1 shows an example of a sample that was considered to have poor Per Sequence GC Content and poor Per Base Sequence Content. A lot of our samples also contained overrepresented sequences. Some of these overrepresented sequences were identified as TruSeq Adapters, while a few others were found to be Illumina Multiplexing PCR Primers or Illumina Single End PCR Primers.

Furthermore, many of our overrepresented sequences were long strings of Gs. These poly-G strings were most present in our negative control samples, where instead of swabbing metal, we exposed the swab to the air. Our negative control samples did not pick up as much DNA as our experimental samples, and it was therefore harder for the Weill Cornell Cluster to differentiate between experimental DNA and poly-A and G string adapters.

This could potentially have also affected the integrity of our per base sequencing content and our GC content. Although we did computationally remove adapters and all of our FastQCs showed a satisfactory result for adapter content, it is possible that we were not able to catch all of them due to the limited timeframe.

Samples with four or more unsatisfactory parts in their quality check were removed as well as samples where a significant fraction of the reads were overrepresented sequences such as adapters or long strings of Gs.

Figure 2 is a UMAP that shows how similar samples taken in the morning were to those taken in the evening. The dispersion of the dots in figure 2 for samples taken in the AM was similar to those taken in the PM. This suggests that the microbiome of the subways in the morning was very similar to the microbiome in the afternoon.

Figure 3 is another UMAP that shows to what degree different stations were genetically similar. Although the distribution of all four stations is rather similar, 96th Street seems to be more concentrated on the bottom of our UMAP, while Flushing/Main Street is more prevalent towards the middle and the top. Grand Central and Times Square seem to be dispersed throughout. This data suggests that Flushing and 96th Street, as terminal stations, have slightly more distantly related microbiomes, while Grand Central and Times Square, as transfer stations, have more diverse microbiomes that contain elements similar to both terminals and each other. However, the points are still very dispersed, indicating that there are no strongly present patterns.

Figure 4 shows the average number of species per sample from different stations. On average, samples taken from Times Square had the greatest number of different species, closely followed by samples from Grand Central. Samples from Flushing and 96th Street both had less. This supported our hypothesis that transfer stations like Times Square and Grand Central have a more diverse microbiome. According to the MTA, Times Square is the single busiest station in New York City. The sheer number of different people who pass through every day probably contribute to the diversity we saw in our samples from Times Square.

References

About MetaSUB. MetaSUB. http://metasub.org/about/
Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S., et al. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst. 1, 72–87. https://doi.org/10.1016/j.cels.2015.01.001
Brooks, B., Olm, M.R., Firek, B.A., Baker, R., Thomas, B.C., Morowitz, M.J., and Banfield, J.F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat. Commun. 8, 1814. https://doi.org/10.1038/s41467-017-02018-w
Danko D, Bezdan D, Afshin EE, . . . Mason CE; International MetaSUB Consortium. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021 Jun 24;184(13):3376-3393.e17. doi: 10.1016/j.cell.2021.05.002. Epub 2021 May 26. PMID: 34043940; PMCID: PMC8238498. https://www.cell.com/cell/pdf/S0092-8674(21)00585-7.pdf
Fast QC. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Fresia, P., Antelo, V., Salazar, C. et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. Microbiome 7, 35 (2019). https://doi.org/10.1186/s40168-019-0648-z
Gilbert, J.A., Stephens, B. Microbiology of the built environment. Nat Rev Microbiol 16, 661–670 (2018). https://doi.org/10.1038/s41579-018-0065-5
Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, DuLong C, Baranowski C, Gevers D, Green JL, Morgan XC, Spengler JD, Huttenhower C. Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. mSystems. 2016 Jun 28;1(3):e00018-16. PMID: 27822528; PMCID: PMC5069760. https://doi.org/10.1128/mSystems.00018-16
Hoch, J.M.K., Rhodes, M.E., Shek, K.L., Dinwiddie, D., Hiebert, T.C., Gill, A.S., Salazar Estrada, A.E., Griffin, K.L., Palmer, M.I., and McGuire, K.L. (2019). Soil microbial assemblages are linked to plant community composition and contribute to ecosystem services on urban green roofs. Front. Ecol. Evol. 7, 198, 10.3389/fevo. https://doi.org/10.3389/fevo.2019.00198
Joyner, J.L., Kerwin, J., Deeb, M., Lozefski, G., Paltseva, A., Prithiviraj, B., McLaughlin, J., Cheng, Z., Groffman, P., and Muth, T.R. (2019). Green Infrastructure Design Influences Communities of Urban Soil Bacteria. Front. Microbiol. 10, 982. https://doi.org/10.3389/fmicb.2019.00982
Kang, K., Ni, Y., Li, J., Imamovic, L., Sarkar, C., Kobler, M. D., Heshiki, Y., Zheng, T., Kumari, S., Wong, J., Archna, A., Wong, C., Dingle, C., Denizen, S., Baker, D. M., Sommer, M., Webster, C. J., & Panagiotou, G. (2018). The Environmental Exposures and Inner- and Intercity Traffic Flows of the Metro System May Contribute to the Skin Microbiome and Resistome. Cell reports, 24(5), 1190–1202.e5. https://doi.org/10.1016/j.celrep.2018.06.109
Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., Rovira, P., Abdo, Z., Jones, K. L., Ruiz, J., Belk, K. E., Morley, P. S., & Boucher, C. (2017). MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic acids research, 45(D1), D574–D580. https://doi.org/10.1093/nar/gkw1009
Leung, M.H., Wilkins, D., Li, E.K., Kong, F.K., and Lee, P.K. (2014). Indoor-air microbiome in an urban subway network: diversity and dynamics. Appl. Environ. Microbiol. 80, 6760–6770. https://doi.org/10.1128/AEM.02244-14
Lax, S., Sangwan, N., Smith, D., Larsen, P., Handley, K.M., Richardson, M., Guyton, K., Krezalek, M., Shogan, B.D., Defazio, J., et al. (2017). Bacterial colonization and succession in a newly opened hospital. Sci. Transl. Med. 9. https://doi.org/10.1126/scitranslmed.aah6500
Maritz, J.M., Ten Eyck, T.A., Elizabeth Alter, S., and Carlton, J.M. (2019). Patterns of protist diversity associated with raw sewage in New York City. ISME J. 13, 2750–2763. Martı´nez, J.L. (2008). Antibiotics and antibiotic resistance genes in natural environments. Science 321, 365–367. https://doi.org/10.1038/s41396-019-0467-z
MetaSUB International Consortium (2016). The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome 4, 24. https://doi.org/10.1186/s40168-016-0168-z
New England Biolabs (2020). NEBNext Ultra II DNA Library Prep Kit for Illumina.x
Ryon, Krista (2022). MetaSUB: Environmental DNA/RNA Sampling. Mason Lab
Subway and bus ridership for 2020. MTA, 2020. https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2020
Wood, Derek (2020). About Kraken 2. https://github.com/DerrickWood/kraken2/wiki/About-Kraken-2
Zhu, Y.-G., Gillings, M., Simonet, P., Stekel, D., Banwart, S., and Penuelas, J. (2017). Microbial mass movements. Science 357, 1099–1100. II OPEN ACCESS Cell 184, 3376–3393, June 24, 2021 339. https://www.doi.org/10.1126/science.aao3007
ZymoBIOMICS (2021). ZymoBIOMICS MagBead DNA Extraction. https://files.zymoresearch.com/protocols/_r2135_r2136_zymobiomics_magbead_dna-rna.pdf

Discussion

We concluded that our expectations that transfer stations would show more genetic diversity were supported. Both the greater average number of species per sample and distribution on the UMAPs suggest that transfer stations showed greater variety in their microbiomes.

However, our sample set was rather limited and, while there were patterns in our data, it is uncertain how significant they are. Sequencing data was displayed in the form of UMAPs and bar graphs in order to visualize the trends we observed. Had we had more time, we might have been able to take more samples and conduct more statistical tests to determine how significant our results were. We also did not expect to have to discard so many samples, which may have skewed our results a little. We had to adapt our data analysis to the new sample set.

Furthermore, we did not expect to see so many adapters in our sequencing data after we performed the quality checks. It is possible that we either made a mistake in our sample cleanup protocol, or we did not successfully computationally remove adapters. If this study were to be repeated we might use a different library prep protocol.

This study raises questions about whether or not other stations and surfaces will follow the same trends in microbial diversity. This study included only four stations because of time restraints, so it is difficult to form larger conclusions from such limited data. It might also be beneficial to repeat this experiment on other mass transit systems in other major urban centers to see if these results are a common phenomenon. If we had more time, we would have also liked to take a closer look at some of our data and perform further computational analyses, or compared other factors such as surface material or day of the week.

Figure 1A

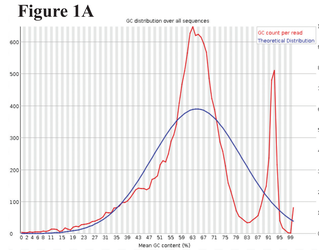

Figure 1B

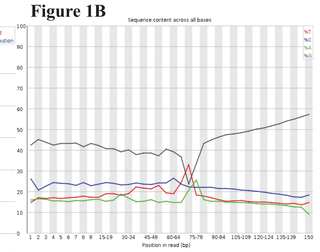

Figure 2

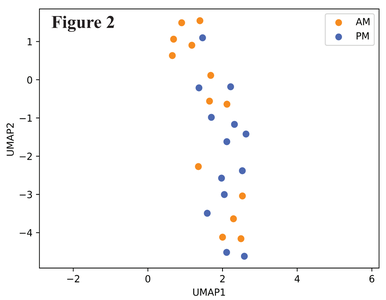

Figure 3

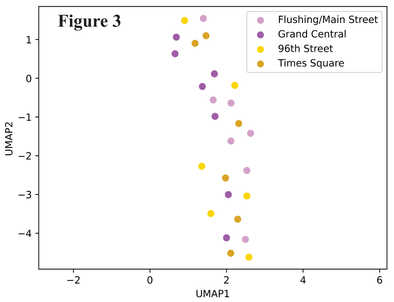

Figure 4

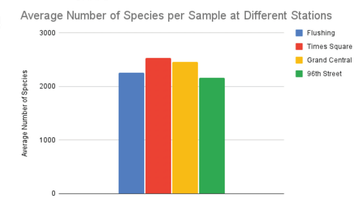

Figure 1: Sample 20.2 – Per Sequence GC Content and Per Base Sequence Content (an example of our poor GC content samples)
Figure 2 - A UMAP comparison between samples taken in the morning and the afternoon showed little difference.
Figure 3 - A UMAP comparison between samples taken at different stations suggested that terminal stations were more genetically distinct
Figure 4 - A bar graph showing the average number of species per sample at each station indicates that, on average, more species were found in transfer stations than in terminal stations.
Figure 5: A flowchart depicting our entire process, from start to finish.

Figure 5