Assessing reproducibility of metagenomics studies and diversity of public transport systems microbiome profiles of New York, Boston and Sacramento cities

Alina Frolova

a.o.frolova@imbg.org.ua, Institute of molecular biology and genetics Kyiv, Ukraine

Introduction

As a new members of MetaSUB Consortium we were greatly interested in analyzing Boston (Hsu et al., 2016), New York (Afshinnekoo et al., 2015) and Sacramento cities microbiome profiles to point out important issues and problems before upcoming global City Sampling Day 2017. We aimed to:

- perform detailed quality control of raw sequences
- evaluate collected metadata in the context of creating uniform specification for data collection
- assess reproducibility of OTU abundances calculation
- verify Yersinia pestis and Bacillus anthracis presence in NY microbiome profile
- investigate biodiversity vs biolocation

Data and Methods

We considered samples sequenced with Illumina HiSeq only, i.e. omitted 16S and MiSeq data. Due to limited computational and storage resources we are still processing NY samples, here we present a subset uniformly distributed among surface material types and date of collection (including positive controls and mixed sample). In total we analyzed 84 samples: Boston = 25, NY = 41, Sacramento = 18.

Raw fastq files were analyzed with FastQC (Andrews et al., 2010) and results were aggregated using MultiQC (Ewels et al., 2016), we repeated this procedure each time we applied any modification to the data to check the impact of the processing. All the data was processed with AfterQC (Chen et al., 2017) tool, which does automatic filtering, trimming, error removing and quality Control for fastq data. Nevetheless, Boston and NY datasets required additional adapters trimming, so we used Trim Galore (Krueger, 2015) to remove Nextera adapters, optionally we used Prinseq (Schmieder et al., 2011) just to filter reads by length (we eliminated reads < 70 bp).

Human DNA was removed with KneadData (Hsu et al., 2016) and abundances were calculated on the species level with Clark (Ounit et al., 2015). R metagenomeSeq (Paulson et al., 2013) package was used for differential abundances testing and heatmaps plots producing. PCA plots were generated with ggplot2 (Wickham, 2016) R package and interactive circular microbiome profiles were created with Krona (Ondov et al., 2011).

All the scripts and tools launch arguments were submitted to github repository, plots and any other supplementary data will be submitted after CAMDA committee approval.

Results

Quality control

Sacramento data has good initial raw sequences quality, so we processed it with AfterQC only. NY and Boston data had quality drop closer to the end of the reads and ambiguous adapters content, which is most pronounced in Boston data that has two quality drop gaps in addition (Fig. 1). We tried to understand the possible cause of adapters overrepresentation, which is described in the github issue we created. The most plausible conclusion is that when using Illumina NexteraXT transposon protocol the resulting insert size is highly sensitive to the concentration of DNA used, and some fragments may have an insert shorter than the length of a single read, which results in the presence of adapter at the end of the read (Turner, 2014).

Importantly, original papers reported only quality trimming ($Q > 20$) without mentioning adapters problem, therefore even after filtering by length we obtained more reads that could be mapped properly.

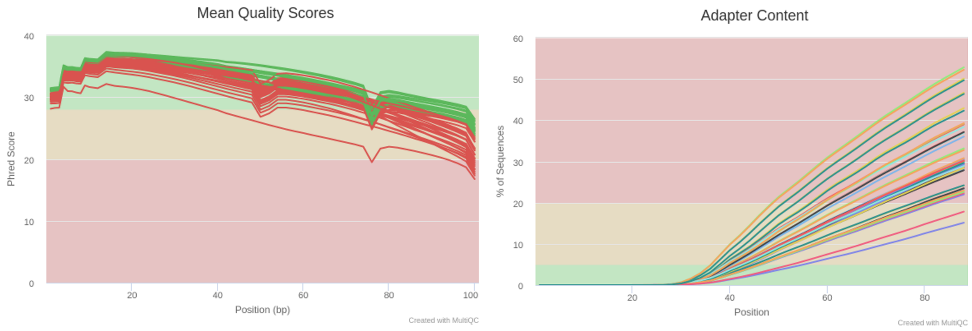

Figure 1: Boston data quality report: mean quality score per base and cumulative percentage count of the proportion of the library which has seen each of the adapter sequences at each position.

Table 1: 32 metadata variables merged from three studies and FastQC reports. Each city dataset contains only subset of these variables.

ID	hs_dna	avg_seq_length	dups	GC	Mseqs	load_date
Mbases	Mbytes	release_date	sample_name	sampling_place	lane	surface_material
no_swabs	precip_rain	rel_humidity_avg	surface_type	station	line	avg_air_temp_F
collection_date	city	latitute	longitude	avg_weekly_riders	borough	ground_level
avg_abs_humidity	avg_dew_point	traffic	unknown_species			

Merging metadata

Diverse metadata was provided per each city for the CAMDA challenge, nevertheless, we noted that original papers describing NY and Boston microbial community study contain more data in fact. The most striking example is an absence of surface type (bench, turnstile etc) in NY metadata sheet, which we extracted from supplementary material of the paper. We also incorporated statistics derived from fastq files such as GC content or average reads length (Table 1). There is a need of a clear specification standard for metagenomics metadata to avoid misinterpretation (e.g. it should be stated if air temperature was measured inside or outsider the station) and ensure presence of important variables (such as person who swabbed the sample as results could depend on that person protocol misinterpretation). The complete summary of metadata issues will be provided to MetaSUB group, responsible for standard implementation.

OTU abundances

We identified the correlation of the top abundant species we found and the species reported in NY and Boston papers. Thus, four species from Table 2 are present in NY top taxa, while Cutibacterium (formerly Propionibacterium) acnes, Micrococcus luteus and Staphylococcus epidermidis are common for Boston and our analysis.

Describing Sacramento profile, Variovorax paradoxus was recently identified as a member of methylotrophic community in the human oral cavity (Anesti et al., 2005), Alteromonas mediterranea for the most part found in the deep water column of the Mediterranean Sea (Ivanova et al., 2015), and Modestobacter marinus (strain BC501) is a bacterium isolated from deep-sea sediment collected from the Atlantic Ocean (Xiao et al., 2011). Interestingly, the relative amount of viruses in Sacramento data is bigger then in NY or Boston, in some samples viruses account for 6% of identified species. Most dominant viruses are Cotesia congregata bracovirus, which has one of the largest genomes known of any virus (567,670 base pairs), and Glypta fumiferanae ichnovirus. For both viruses parasitoid wasp serves as host.

The authors of NY study reported observing Yersinia pestis (the causative agent of plague) and Bacillus anthracis (the causative agent of anthrax) as part of the "normal subway microbiome", while authors of Boston paper did not find these species after re-analyzing NY data with more strict threshold (taxa with at least 0.1% abundance in at least 1% of samples) (Gonzalez et al., 2016). In our study we found Yersinia pestis in sample SRR1749429 with 0.0229636% abundance and Bacillus anthracis in sample SRR1749465 with 0.032334% abundance (the percentage here calculated based on all the reads, including unknown ones). So, it is very important to define plausible threshold, especially in case of pathogenic agents, to avoid different interpretation of the same data. Additionally, since bacterial load vary across the samples greatly it may be beneficial to reveal alterations in absolute abundance of specific OTUs by using spike-in bacteria added to each sample before sequencing (Stämmler et al., 2016).

There are also differences between original papers and our results. For example, among top species in NY data we identified Serratia marcescens - an opportunistic pathogen whose clinical significance has been appreciated only in the

Table 2: Top 5 OTU abundances (average across the samples)

Boston		New York		Sacramento	
Species	Count (%)	Species	Count (%)	Species	Count (%)
UNKNOWN	83.371	UNKNOWN	48.495	UNKNOWN	96.160
Cutibacterium acnes	8.688	Pseudomonas stutzeri	20.475	Alteromonas mediterranea	0.300
Micrococcus luteus	0.415	Stenotrophomonas maltophilia	5.794	Variovorax paradoxus	0.241
Staphylococcus epidermidis	0.308	Serratia marcescens	2.088	Modestobacter marinus	0.235
Modestobacter marinus	0.233	Enterobacter cloacae	1.982	Geodermatophilus obscurus	0.108
Streptococcus sanguinis	0.184	Bacillus cereus	1.868	Nocardioides sp. JS614	0.092

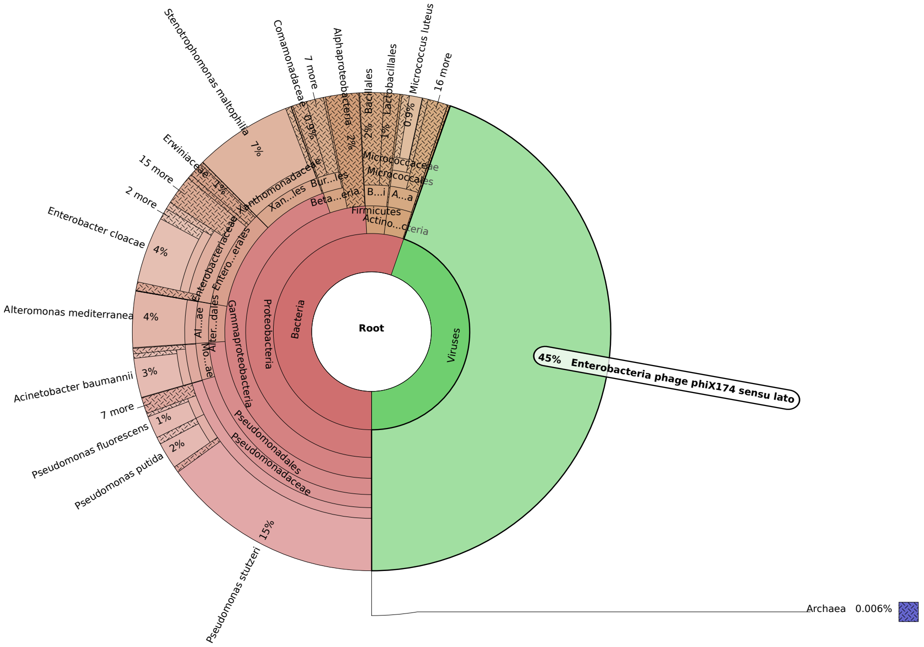

Figure 2: SRR1750076 sample from NY study contaminated by Enterobacteria phage phiX174 sensu lato.

last four decades. While S.marcescens is a rare cause of community-acquired infections, it has emerged as an important nosocomial healthcare-associated pathogen and a frequent source of outbreaks of hospital infection (Merkier et al., 2013).

Contrast to NY findings our analysis reported zero or less then 0.01% on average presence of Enterobacteria phage phiX174 sensu lato, except one sample. SRR1750076 represents pool of samples and has 45% of phiX among indentified species (see Fig. 2, or download and launch with browser interacive Krona plot). It might indicate contamination due to phiX174 being used as control frequently during Illumina sequencing runs (Mukherjee et al., 2015). But to support the hypothesis we need to finish processing NY samples and additionally verify pooled sample preparation and sequencing details with study authors as we did not find any mentions of such phiX prevalence in that particular sample. We aim to filter all the kitome before deriving final conclusion about NY and other cities microbial communities.

Biodivercity vs biolocation

Analyzing Boston data we confirmed that surface material was a major determinant of samples microbial community composition. Nevertheless, the sampling is biased towards surface type as almost all the materials correspond to one surface. For example, only grips were made from PVC material, while the latter has the strongest connection with large Human DNA amount with 56% on average and standard deviation of 9%. The future studies should focus on more rich sampling in the context of different meta-features.

NY data does not have clear association with one type of variable. Having in mind the diversity of sampling we agree with Boston study authors, that additional information is required to infer true correlations, such as details on exact sampling protocol, e.g. which parts of objects were swabbed.

Table 3: Sacramento. Potentially paired samples.

Paired	Odd	Possible Cause
1B, 1A	1C	1B and 1A have smallest amount of unknown species
5A, 5C	5B	5B has 35.88% of Human DNA
3B, 3C	3A	Station had many people when sampling 3A
4B, 4A	4C	Rained 2 mins right before sample 4C was taken
6A, 6B	6C	6C has 53.11% of Human dna; 6A, 6B - the bench and the machine were covered

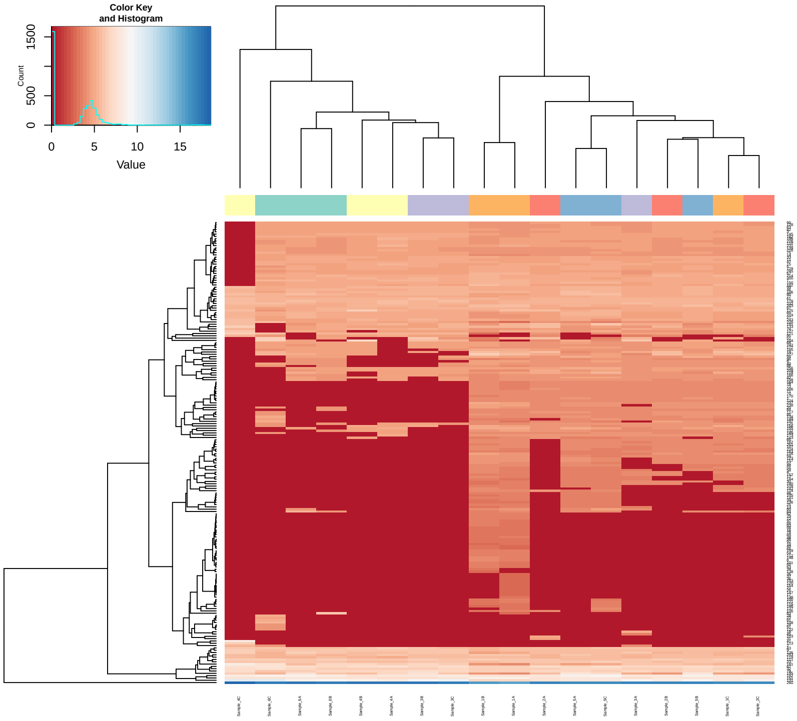

Figure 3: Sacramento OTUs heatmap. Groups colors represent different stations

In the Sacramento study the sampling was done on the above ground stations under the condition of strong wind. Since the data was sequenced in Mason's lab and swabbing protocol is very similar to NY study, we expected small amount of Human DNA contamination, however sample 5B (ticket machine) has 35.88% and 6C (platform rail) has 53.11% of Human DNA, while the rest of the samples - around 1%. PCA plot (Fig. 4) shows that those two samples are clustered apart, but before making any biological conclusion we have to verify if there is a possibility those two samples were accidentally contaminated by the person(s) in charge.

When comparing Sacramento with other studies it has the highest number of unidentified species (Table 2) - 96% on average, with samples 1A and 1B having 90% and 91% and clustered apart of the rest samples on the PCA plot (Fig. 4). Since the study is unpublished it is unknown if the wet-lab procedures could have contributed to such difference.

The heatmap of Sacramento OTU abundances (Fig. 3) showed interesting pattern of two samples from one station forming a group (for other meta-variables we did not find plausible hierarchical clustering). We carefully studied provided and derived metadata to summarize possible cause of such behavior in Table 3. Additionally, we performed differential abundance testing of samples collected from 8th & O rail station with the split platforms located on each side of 8th Street, thus having East and West sides. Sample 4C from East was collected right after the rain (it has the smallest number of OTUs on Fig. 3), while sample 3A was collected during intensive human traffic. Table 4 shows that 8th & O (east) has decrease in species with Cutibacterium acnes on the top, commonly found on sebaceous skin sites, which comprise the chest, back, and face.

So, we think that only after performing larger and more uniform study to sample stations under different weather conditions (especially important for open areas) and human traffic researchers can verify if the microbiome profiles are really station-specific in case of Sacramento city.

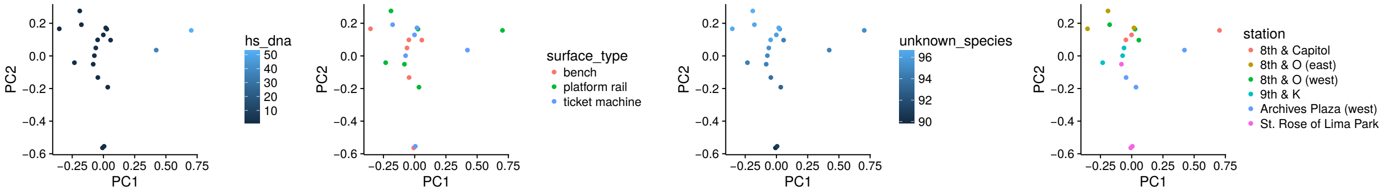

Figure 4: Sacramento species abundances PCA plots: Human DNA %, surface type, unknown species %, stations.

Table 4: Differentail abundance testing. Sacramento: 8th & O (west) vs. 8th & O (east) station, top 5 species.

logFC	AveExpr	adj.P.Val	Species
-12.11937	6.794665	2.882703e-02	Cutibacterium acnes
-11.01962	8.700627	2.396800e-10	Polaromonas naphthalenivorans
-10.99391	8.317716	8.530000e-12	Methylobacterium populi
-10.88727	6.445794	1.120400e-09	Streptomyces coelicolor
-10.78604	7.075633	7.429700e-10	Streptomyces cattleya

Conclusion

We managed to reproduce major number of top OTU abundances of NY and Boston cities microbiome profiles. We confirmed that all three cities harbor distinct microbial communities. Nevertheless, we identified that all three studies are biased towards the selection of particular metadata, sampling plan and swabbing plans, which introduces a lot of hidden variables and makes the comparison more difficult.

We agree with the authors of Boston study that it is important to sample fewer, more controlled environments with greater specificity and uniform coverage of meta-variables. We plan to extend this work further by studying antimicrobial resistance markers and filtered Human DNA.

Acknowledgments

We acknowledge the access to the computing resources of Chair of Bioinformatics Research Group of Boku University Vienna, Austria, and we are thankful to Paweł P. Łabaj and Maciej Kańduła for technical assistance. We would also like to express out gratitude to Stefan Tsonev and Dimitar Vassilev from Sofia University, Bulgaria, for fruitful discussions.

References

Afshinnekoo et al. (2015), *Cell systems*, 1(1): 72–87.
Andrews et al. (2010), *https://www.bioinformatics.babraham.ac.uk*.
Anesti et al. (2005), *Environmental Microbiology*, 7(8): 1227–1238.
Chen et al. (2017), *BMC bioinformatics*, 18(3): 80.
Ewels et al. (2016), *Bioinformatics*, 32(19): 3047–3048.
Gonzalez et al. (2016), *mSystems*, 1(3): e00050–16.
Hsu et al. (2016), *mSystems*, 1(3): e00018–16.
Ivanova et al. (2015), *Antonie Van Leeuwenhoek*, 107(1): 119–132.
Krueger (2015), *https://www.bioinformatics.babraham.ac.uk*.
Merkier et al. (2013), *Journal of clinical microbiology*, 51(7): 2295–2302.
Mukherjee et al. (2015), *Standards in genomic sciences*, 10(1): 18.
Ondov et al. (2011), *BMC bioinformatics*, 12(1): 385.
Ounit et al. (2015), *BMC genomics*, 16(1): 236.
Paulson et al. (2013), *Nature methods*, 10(12): 1200–1202.
Schmieder et al. (2011), *Bioinformatics*, 27(6): 863–864.
Stämmler et al. (2016), *Microbiome*, 4(1): 28.
Turner (2014), *Frontiers in genetics*, 5: 5.
Wickham (2016), *Ggplot2: Elegant Graphics for Data Analysis*. Springer.
Xiao et al. (2011), *International journal of systematic and evolutionary microbiology*, 61(7): 1710–1714.