Strain-level bacterial and viral diversity in the MetaSUB dataset

M. Zolfo*, F. Pinto, F. Asnicar, F. Beghini, P. Manghi, E. Pasolli, S. Manara, A. Tett and N. Segata

Laboratory of Computational Metagenomics (CIBIO)
* Presenting author

The study of the microbiomes associated with the urban and built environment is of increasing research interest, as it is recognized that the microbial diversity associated with the places where people live and work everyday influences human health. While endeavours like the MetaSUB [1] Consortium have started to characterize the structure and the composition of the microbiome of our cities, improved resolution in metagenome profiling is needed to uncover key microbial features that are distinguishing different strains within the same species. Small genetic differences among microbes of the same species can characterize ecological niches associated with different places, and little variations in the genomic landscape of microbes can convey important phenotypical variations. For example, the presence of toxins, antibiotic resistance genes, and virulence factors is at the base of the variable level of pathogenicity associated with bacterial species like *Escherichia coli* [2] and Staphylococcus aureus [3], that can be useful commensal or life-threatening pathogens. However, metagenomic profiling failed so far to provide strain-level characterization of the microbial diversity on large datasets, preventing the use of metagenomics as a tool for strain-level microbial epidemiology and population genomics.

We analysed here the 1614 metagenomes of the MetaSUB/CAMDA2017 dataset with a set of recently developed[4,5] and new computational tools to unravel the strain-level diversity in the public transportation microbiome All microbial strain profilers were applied directly to the the raw reads and are applied on non-human microbiomes for the first time in this study. We first applied MetaMLST[5], a computational tool that profiles the strains in metagenomes using the Multi Locus Sequence Typing approach[6], and is capable of identifying both known and previously unseen (novel) sequence types (STs). Once profiled, such STs allow tracing (potentially pathogenic) strains across samples and comparing the identified strains with the extensive set of STs deposited in publicly available MLST databases containing more STs than available reference genomes. In our analysis, a total of 109 species were profiled, and 642 STs were observed. Of those, 500 are being observed here for the first time and differ by at most 10 SNVs from their closest known reference ST (**Figure 1A**). Among these, the most prevalent species were *Acinetobacter baumannii*, *Enterobacter cloacae* and *Stenotrophomonas maltophilia*: all species that are both known to be abundant in the environment but that are also associated to potential pathogenic infections in humans. Using *E. coli* profiles as an example, STs can be epidemiologically modelled using Minimum Spanning Trees which highlighted that 10 of the 19 identified *E. coli* STs are very closely related and within the *E. coli* phylotype A. The remaining STs are instead clearly distinct clones from phylotypes often populating the human gut microbiome, that would be consistent with the hypothesis of a human origin of these strains (**Figure 1B**).

To further extend the strain-level profiling to larger portions of the genome, we applied StrainPhlAn [4], a tool that characterizes strains by analysing single nucleotide variations

---

[1] *The MetaSUB International Consortium, Microbiome, 2016*
[2] *NJ Loman et al., Jama, 2013*
[3] *Wielders et al., the Lancet, 2001*
[4] *Truong et al., Genome Research, 2017*
[5] *Zolfo et al., Nucleic Acid Research, 2016*
[6] *Maiden et al., PNAS, 1998*

(SNVs) in clade-specific marker genes[7]. We reconstructed whole-genome scale phylogenies for the strains belonging to the six most abundant species in the dataset (*Pseudomonas stutzeri, S. maltophilia, Enterobacter cloacae, Propionibacterium acnes, Acinetobacter pittii and Acinetobacter radioresistens*), and found a remarkable sub-species variability in the samples. As expected, when associating such strain diversity with location and surface type we did not found strong genetic patterns, consistently with the high level of microbial seeding these samples are exposed to. Nevertheless, we were able to identify very closely related strains present along the same subway line (trains and stations) in the Boston dataset (**Figure 2A**). In other cases, we could identify the same clones within the same station, such as for the *P. stutzeri* strains found on lines NY-M and NY-R (data not shown).

The genetic variability across samples of the profiled strains highlighted in some cases the presence of discrete clusters (sub-species) within species. Specifically, we identified two and three subspecies-level clusters for *P. stutzeri* and S*. maltophilia* respectively (**Figure 2D,2E**). Even though no specific associations with sample types were found, these genetic niches could reflect different functional potential properties that are adapting to similar environments.

Another key aspect in microbial ecology is the presence of multiple related strains of hte same species in a sample. By analysing the polymorphisms in clade-specific marker genes for all the considered species, it is possible to infer the presence of more than one strain of a given species. We performed such analysis with StrainPhlAn, and we highlight that *E. cloacae*, *P. acnes* and *A. pittii* have high rates of polymorphism, suggesting that in the large majority of samples these species are represented by multiple strains (**Figure 2C**). These polymorphic rates tend to be higher than those found in the human microbiome [4] probably owin to the higher exposure to diverse strains and lower strain adaptation costs of environmental microbiomes.

Finally, we developed and applied a new computational pipeline aimed at profiling the viral fraction of microbial communities. This is based on mapping the reads against known or newly assembled phage genomes and the identification of samples-specific SNVs for each detected viral genome. We identified a total of 73 viral genomes present in at least 5 samples. By reconstructing the dominant allele on the identified SNV positions, we then reconstructed the sample-specific genomes of several members of the virome. For example, the pipeline reconstructed the genomes of 11 distinct bacteriophages usually associated to *Bacillus* species with a variable level of similarity with previously available sequences (**Figure 1B**). More in depth analysis on the viral fraction of the MetaSub dataset and their relation with the bacterial community are being conducted, and will be ready to be presented at CAMDA/ECCB 2017.

Together, these findings show that it is possible to analyse and trace the microbes of the urban environment at strain-level, directly from the raw reads and without the need for assembly. Coupled with ongoing assembly-based efforts to uncover genomes of novel and uncharacterized microbial species, we show the potential use of metagenomic data for large-scale epidemiology, bio-surveillance, outbreak tracking, and bacterial-resistance profiling.

[7] *Segata et al., Nature Methods, 2012*

A

	Total STs	Known STs	New STs
A. baumannii	192	69	123
E. cloacae	152	63	89
S. maltophilia	113	15	98
C. ronobacter	68	2	66
S. enterica	58	0	58
K. pneumoniae	51	13	38
B. cereus	32	15	17
K. oxytoca	30	12	18
A. chromobacter	29	2	27
E. faecalis	24	6	18
S. suis	24	0	24
P. acnes	22	4	18
E. coli	19	3	16
P. fluorescens	8	1	7
P. aeruginosa	6	5	1
C. botulinum	6	1	5
TOTAL	**642**	**142**	**500**

B

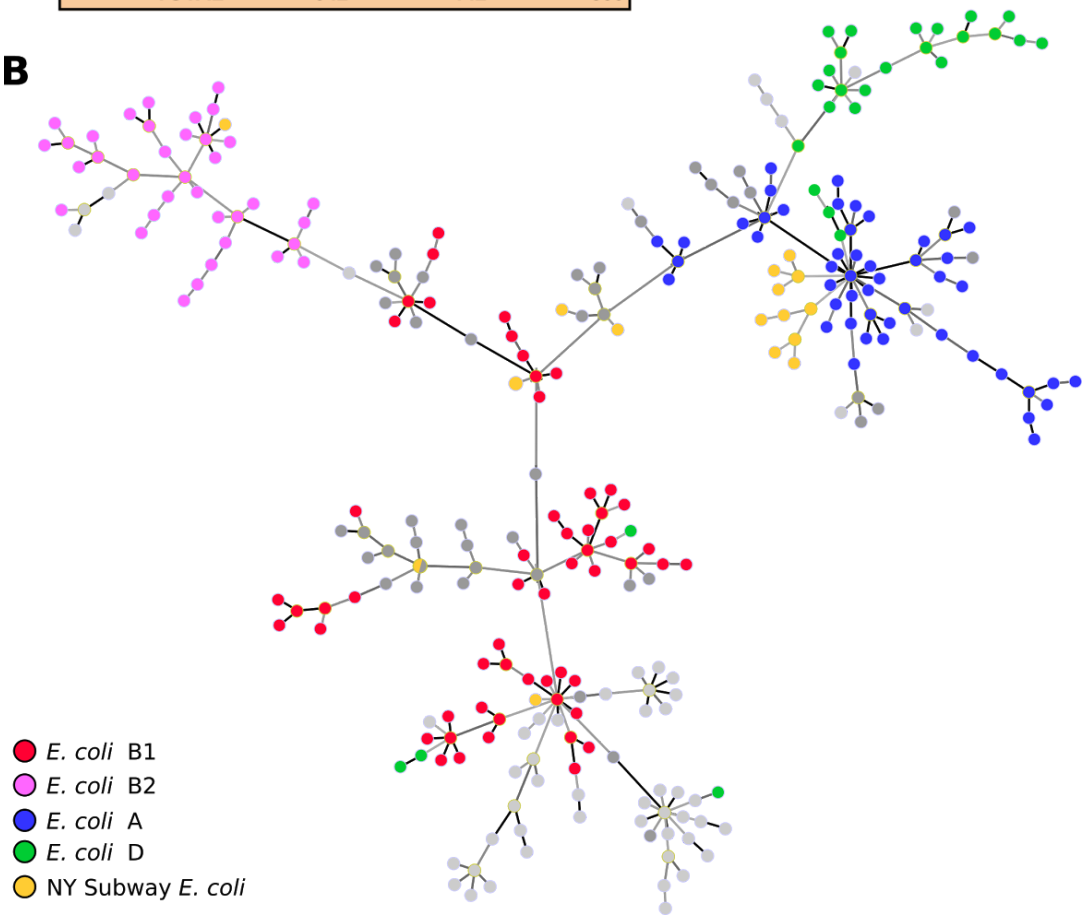

Figure 1: Multilocus Sequence Typing characterisation of bacteria in the New York subway. A) *Detected Sequence Types (STs), by species. MetaMLST detects calls for a new STs when at least one different nucleotide is detected in the overall profile, and at most 10 SNVs are present with respect to the closest known reference ST.* ***B)*** *Minimum Spanning Tree analysis performed using PHYLOViZ [8] with the goeBURST Full-MST algorithm. The nodes are the E. coli STs of the New York Subway (in yellow) and the known STs stored in public repositories(other colours) [9].*

[8] A. P. Francisco et al, BMC Bioinformatics, 2012
[9] Jolley & Maiden, BMC Bioinformatics, 2010

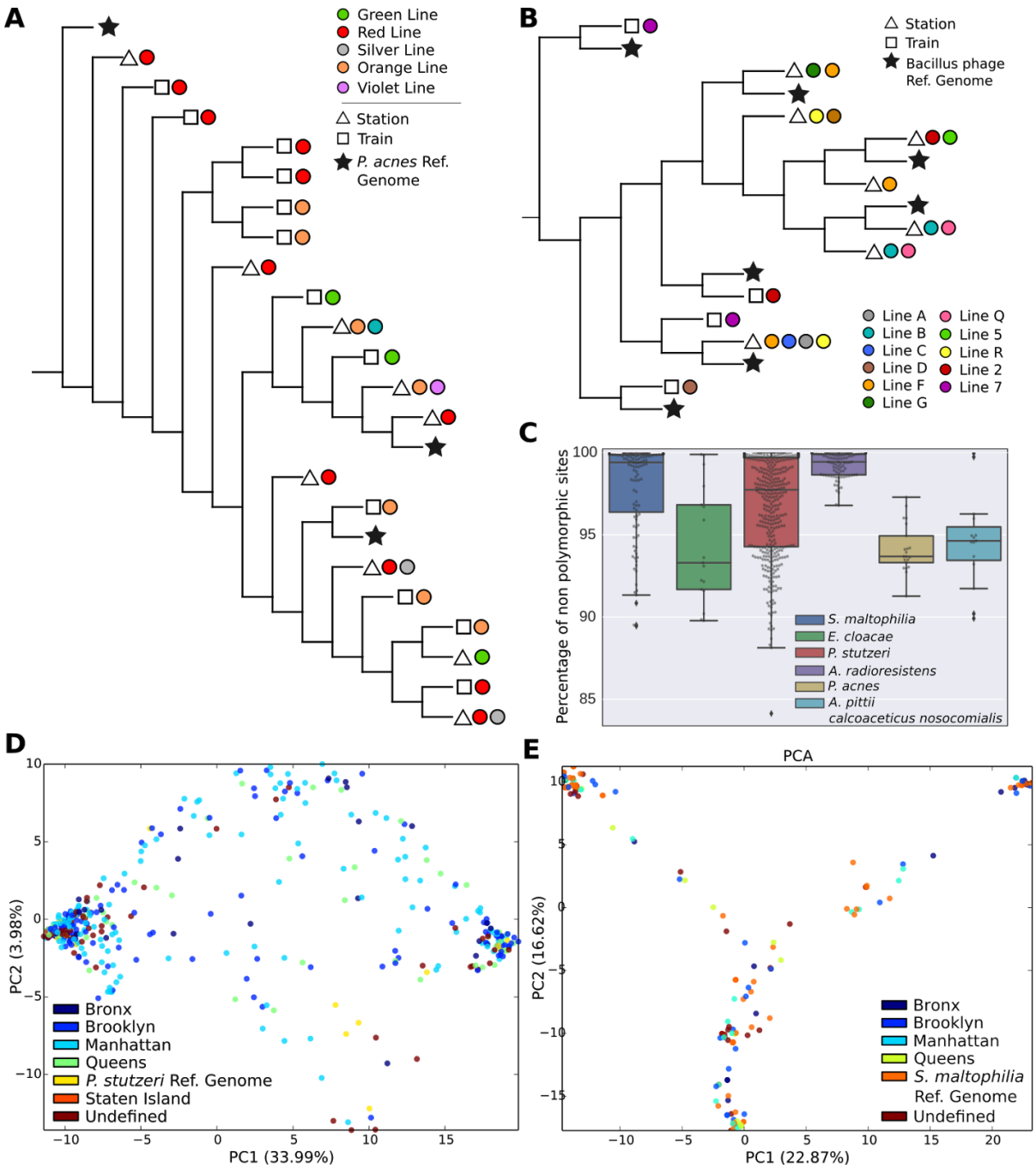

***Figure 2: Phylogenetic analysis of bacteria and phages in the MetaSUB dataset**. **A)** Phylogenetic tree of P. acnes strains recovered from the metagenomic samples. The tree is built on the concatenated highly informative positions of StrainPhlAn markers. Coloured circles represent subway lines. Station connecting more lines have more than one color. Markers extracted from reference genomes are indicated with stars **B)** Phylogenetic tree of Bacillus phages built on the reconstructed whole genomes of each virus. Colors and markers describe lines and trains/stations in **A)**. The trees were realised with RAxML v8 [10] using the GTRCAT model. **C)** Percentage of non-polymorphic sites in the reconstructed StrainPhlAn markers. **D) E)**: PCA analysis of the reconstructed StrainPhlAn markers for the P. stutzeri and S. maltophilia species.*

[10] *A. Stamatakis et al, Bioinformatics, 2014*