Assessment of urban microbiome assemblies with the help of targeted mock communities

Samuel M. Gerner[1,3], Josef W. Moser[2], Thomas Rattei[3], and Alexandra B. Graf[1,*]

[1] University of Applied Sciences FH Campus Wien, Department Bioengineering, Vienna, Austria

[2] Austrian Centre of Industrial Biotechnology (ACIB), Vienna, Austria

[3] Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

*Corresponding author

Background

Urban microbiomes differ from other known microbiomes in their comparatively high population dynamics, especially when considering areas with a high fluctuation of bypassing humans as subway/railway stations of public transport systems (~238 million trips/year in Boston, ~14 million trips/year in Sacramento). The MetaSUB International Consortium (The MetaSUB International Consortium 2016) aims to improve quality of living, city utilization and planning through the detection, measurement and design of metagenomics studies within urban environments. Several cities have already published results of their urban microbiomes, like New York City and Boston (Afshinnekoo et al., 2015; Hsu et al., 2016), giving insights into the bacterial and human diversity showing that bacteria found in their samples mainly represent harmless species, but also that a large part of the samples contains still unknown DNA. To detect novel species and to enable a detailed analysis of microbe-microbe communities or host-microbe interactions, metagenomic reads have to be assembled into ideally complete genomes. However, to our current knowledge, no other study tried to accomplish assemblies of urban microbiomes so far. Assembly quality and genome binning approaches are influenced by a wide range of factors. These Influences affect computational performance, detection of low abundant taxons and species,well as the purity of the bins from said assemblies to resolve bacterial genoms at strain-level. To help people dealing with a plethora of assembly tools, it is essential to provide clear assessment parameters and quality measures for assembly methods. In Sczyrba et al. (2017) different methods and strategies were compared to help in obtaining good and contamination-free bin sequences. They found that assembly tools perform very differently, depending on the features of the metagenome sample. These features include population diversity, sequencing quality, sequencing depth and input material. High community diversity, especially the presence of closely related microbial strains, can decrease assembly performance dramatically and is one of the main challenges in metagenomics analysis.

Aim of the study

- Create **assemblies** from typical urban metagenome datasets and assess the quality of these assemblies.
- Create a **mock community** with **typical features of an urban metagenome dataset**. Analyse the performance of assemblers using the created mock community and propose a set of **recommendations** in reference to **sequencing parameters to increase assembly and binning quality** of urban metagenome data.

In this study, we have constructed multiple metagenome assemblies of Sacramento bench samples from five different stations (Samples 1-4A, 6A) by assembling single samples as well as established an assembly of all bench samples were pooled together. Pooling the samples should give us more power to analyse low abundant organisms in the dataset. We used these datasets to provide first results, and also plan to apply the approaches to all provided shotgun sequenced metagenome datasets in the CAMDA MetaSUB Inter-City Challenge.

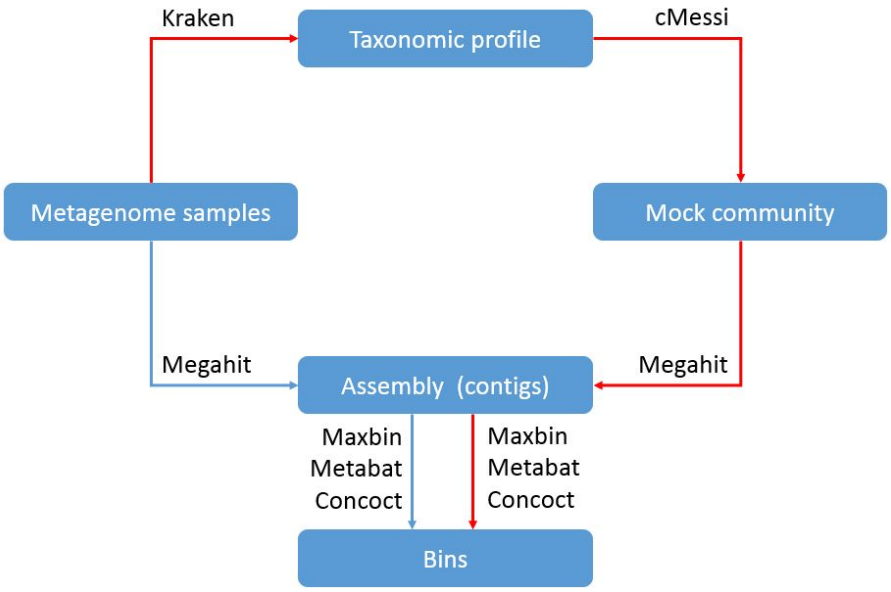

Figure 1. Schematic overview of the approach used to analyse the five samples from the Sacramento Bench dataset using tools for establishing the Mock community and for assembling the metagenomes with various different tools.

To further assess features of urban metagenomes that influence assembly performance, we created mock communities based on taxonomic profiles of the reads from the original samples. These mock communities are generated from the assigned taxa, representing the population diversity of urban metagenome samples. All included taxa are based on taxonomic profiles from Kraken and their respective reference genomes obtained from the NCBI Reference Sequence Database. Mock communities created from simulated reads with high sequencing quality, varying sequencing error rates, varying sequencing depths, varying levels of contamination, as well as varying levels of strain diversity are evaluated. These mock communities will be used to

determine optimal sequencing parameters in order to provide recommendations for an optimal sequencing setup that would enable the creation of high-quality bins out of the assemblies.

Preliminary Results

All Sacramento grey metal bench samples (Samples 1-4A, 6A) were subjected to taxonomic analysis using Kraken (Wood et al., 2014) and MetaPhlAn2 (Truong et al., 2015), which use either a k-mer or marker gene-based approach. Kraken and MetaPhlAn2 were chosen due to good performance in Lindgreen et al. (2016) and Sczyrba et al. (2017). While Kraken could classify 8,176,806 of 156 million quality controlled reads for the Pool of all Sacramento bench samples, MetaPhlAn2 could only classify 138,758 reads due to it's different approach using marker genes. Single samples showed 91-97% unclassified reads using Kraken which indicates at either sequencing artefacts or yet unknown species, both possibilities need to be assessed later on. Relative abundances did not agree fully between both taxonomic methods, with Kraken classifying 52% of all reads as Proteobacteria and 38% as Actinobacteria and MetaPhlAn2 classifying 39 % and 36% as Proteo- and Actinobacteria respectively. Phylums with lower abundances showed even more pronounced differences as Kraken identified 1% Bacteroidetes, 2 % Firmicutes and 3% Cyanobacteria in all reads classified as Bacteria, while MetaPhlAn2 classified 12%, 5% and 1% respectively, showing relative abundances need to be interpreted with caution, as different classification methods can give varying results. Taxonomic profiles of single samples showed a similar pattern.
Unknown species can only be identified by assembling the metagenomic data followed by binning of the resulting contigs. To this end all samples were preprocessed by Trimmomatic (Bolger et al., 2014) and assembled by MEGAHIT (Li et al. 2016). MEGAHIT performed favourably in Sczyrba et al. (2017) as well as in Vollmers et al. (2017). To assess the impact of species diversity without the impact of potential sequencing artefacts, mock communities based on taxonomic profiles of Kraken were created with varying complexity and sequencing depth. Mock community assemblies shown are constituted of all species and their respective subspecies with at least 20,000 reads assigned to species level as well as the presence of a reference genome for the same taxonomic ID in the NCBI Reference Genome Database resulting in 227 genomes selected. Mock communities were created with 10 and 20 million 125 bp long high quality read pairs sampled from the selected reference genomes and assembled in the same manner as the original samples.

Resulting contigs were binned with three different binning programs, namely MetaBAT (Kang et al., 2015), MaxBin (Wu et. al., 2016) and CONCOCT (Alneberg et. al., 2014) due to good performance in Sczyrba et al. (2017), which apply nucleotide composition and abundances to place contigs into genome bins. These bins were also checked for contamination and completeness using single copy genes as provided in CheckM (Parks et al., 2015).

Sample	Assembly size	Contigs	Avg contig size	Max contig	N50	Reads mapping
Sample 1A	101,759,459	113,596	896	49,364	829	17.41%
Sample 2A	71,245,988	91,093	782	46,571	717	24.30%
Sample 3A	58,878,533	78,491	750	100,608	688	18.82%
Sample 4A	87,916,835	118,212	744	48,282	689	27.14%
Sample 6A	71,600,288	94,868	755	46,394	698	21.58%
Pool 12346A	563,529,486	721,126	781	100,559	728	30.82%
Mock 10m	297,100,595	255,456	1163	1,206,229	1216	78,85%
Mock 20m	472,069,039	253,055	1865	1,206,235	2780	93.13%

Table 1. Assemblies statistics of Sacramento grey metal bench samples 1-4A and 6A as well as an assembly of all respective samples together. Mock communities are created as described above, 10m/20m standing for 10/20 million 125 bp long read pairs.

Samples 1-4A and 6A consist of 20-25 million read pairs, of which 5% could be classified as bacterial. Sequencing depth has a major effect on assembly performance. This can be seen by the number of input reads mapping back to the assembly. The alignment rate is highest for the pooled real data sample as well as in the two mock communities, which consist of the respective RefSeq genomes and relative abundances thereof. A two-fold increase in sequencing depth already raises the amount of input reads mapping back to the assembly tremendously.

Sample	Bin	Binner	Completeness	Contamination	Strain heterogeneity	Reads mapping
Pool 12346A	Bin 18	MaxBin	92.26%	7.57%	13.64	0.27%
Pool 12346A	Bin 54	CONCOCT	92.1%	9.69%	69.23	0.08%
Pool 12346A	Bin 32	MaxBin	90.88%	9.66%	68	0.08%
Pool 12346A	Bin 13	CONCOCT	90.52%	5.05%	50	0.09%
Mock 20m	Bin 8	MetaBAT	99.89	0.59	0	5.41%
Mock 20m	Bin 2	MaxBin	99.89	0.65	0	5.52%
Mock 20m	Bin 10	MaxBin	99.78	0.33	0	3.04%
Mock 20m	Bin 25	MaxBin	98.11	10.98	5.88	0.83%

Table 2. Top bins with high coverage of assembled input reads, estimated completeness above 90 % and less than 15% contamination as estimated by CheckM. Input reads are mapped to single bins.

Binning the assembly of the pooled samples resulted in 21, 58 and 113 bins for MetaBAT, MaxBin and CONCOCT respectively, for which 3.96%, 15,00% and 15.44% of all input reads could be mapped to all bins of the respective methods. The same binning for the Mock community with 20 million reads resulted in 27, 92 and 68 bins for MetaBAT, MaxBin and CONCOCT respectively, of which 67.65%, 78.95% and 84.92% of all input reads could be mapped to their respective bins. This demonstrates the possibility to retrieve high quality bins with very low contamination as shown in Table 2 for selected single bins of a community as provided in the selected samples.

To enhance the presented analysis, various types of contamination and their impact will be assessed to propose recommendations in reference to sequencing parameters. This should lead to increased assembly and binning qualities in urban metagenome data. Additionally, bins obtained from the real data samples will be further analysed and refined by reassessing and reassembling high quality bins as well as analysing their functional potential and putative virulence factors. Sequences which could not be classified by taxonomic profilers but grouped together in bins are likely representing unknown species and are thereby an interesting target for further research, whereas contigs which are neither able to be classified nor binned might represent sequencing artifacts. Human contamination was tested, but only about 1,5% of the reads were found to map against the hg38 human reference genome.

References

Afshinnekoo, E.,et al. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Systems*, *1*(1), 72–87.

Alneberg, J.,et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods*, (11), 1144-1146.

Bolger, A. M.,et al. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120.

Hsu, T., et al. (2016). Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. *mSystems*, *1*(3), 1–18.

Kang, D. D., et al. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, (3), e1165

Li, D. et al., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* , 102, pp.3–11.

Mende, D. R., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE*, *7*(2).

Parks, D.,H., et al. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25: 1043–1055.

Sczyrba, A., et al. (2017). Critical Assessment of Metagenome Interpretation – a comprehensive benchmark of computational metagenomics software. *bioRxiv*.

The MetaSUB International Consortium. (2016). The Metagenomics and Metadesign of the Subways and Urban Biomes. *Microbiome*, 24(4): 1–14.

Truong, D. T., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903.

Wood, D. E., et al. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.

Wu, Y. W., et al. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, (32), 605-607.