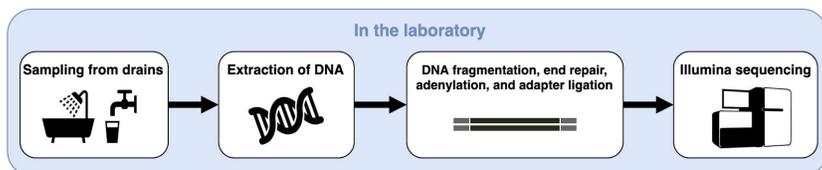


## Introduction

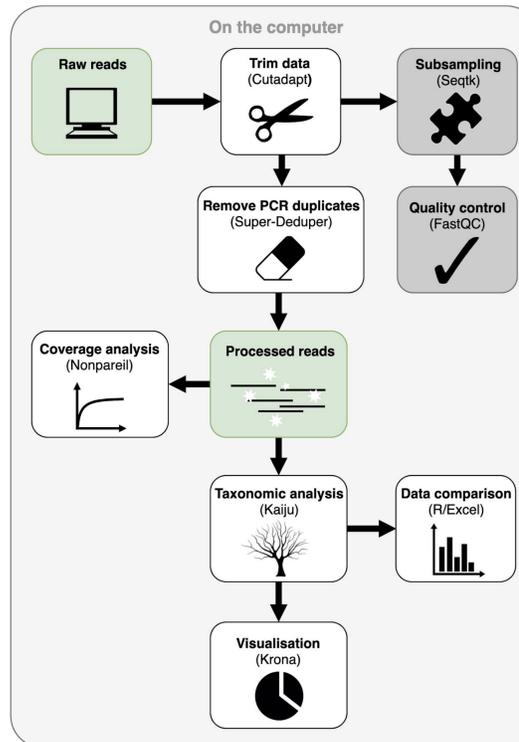
Drains contain a plethora of biofilm producing bacteria, creating micro-ecosystems in practically every bathroom and kitchen around the world. Despite their commonness, much has yet to be learned about the formation, development, and interactions of these complex microbial communities. Many parameters are predicted to influence how bacterial communities are structured. Of these, the life span of the community might be a significant factor. However, to our knowledge, no studies have yet focused on investigating how the age of the plumbing systems affect the bacterial composition of their microbial ecosystems. As such, we wish to utilise metagenomic methods to investigate how the age of the pipes influences the bacterial diversity of three different drain systems: Shower drains, bathroom sink drains, and kitchen sink drains.

## Workflow



### Library creation

A number of samples were obtained from the drains of different kitchens and bathroom sinks and shower drains. The DNA from each sample was first purified and then fragmented. The resulting DNA fragments were end-repaired and adenylated, so that adapters containing labelling barcodes could be attached. Finally, the processed fragments were paired-end sequenced using an Illumina HiSeq2500 system. Of the samples collected, some were eventually discarded due to sequencing related issues, such as primer dimer formation or blank results. The final data displayed in this study thus originated from 32 different samples.



### Fantastic programmes and where to find them

Cutadapt: <https://cutadapt.readthedocs.io/en/stable/>

FastQC: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Super-Deduper: <https://best.github.io/HTStream/>

Nonpareil: <http://enve-omics.ce.gatech.edu/nonpareil/>

Kaiju: <http://kaiju.binf.ku.dk/>

Krona: <https://github.com/marbl/Krona/wiki>

## Data processing and analysis

### Pre-processing and quality control:

In order to prepare sample reads for analysis, it is necessary to perform both quality control and pre-processing. Following programmes were used:

Cutadapt is capable of removing the second adapter sequence from each read (the first were removed in advance). It also removed reads below 30 bp or with a quality score lower than 20. Finally, it was used to trim the first three base pairs from each read, as their quality was generally low based on FastQC analyses.

FastQC can be used to identify read quality (phred score), adapter presence, duplication levels, GC content, and read length. It was here used to examine subsets of each sample before and after Cutadapt processing. These subsets were created with the Seqtk programme.

Super-Deduper is simply used to remove PCR duplicate reads. This prevents disruptive noise due to over-amplification of specific reads.

### Analysis:

After the reads were processed, they were scanned in order to determine their coverage and taxonomic contents. Besides the data management softwares R and Excel, following programmes were used:

Nonpareil estimates the average coverage needed to get a nearly complete metagenome coverage based on the redundancy of reads in metagenomic datasets.

Kaiju taxonomically classifies all sample read pairs. This is done by translating the nucleotide sequences to protein sequences (for all six reading frames), and then finally conducting a backwards Burrows-Wheeler transformation in order to decompress the resulting data. The read pairs were matched to an NCBI database which contains non-redundant proteins from bacteria, archaea, viruses, and eukaryotes (NCBI nr+eu).

Krona is able to visualise Kaiju outputs in diagrams. It also includes the read pair quantities and percentage distributions for each OTU.

## Data quality assessment and analysis

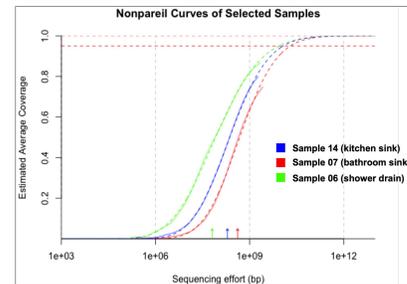


Figure 1: Nonpareil curves for samples 6, 7, and 14. The solid curves indicate the estimated coverage and the dashed curves show the projection curve. The top dashed line indicates 100% coverage while the lower dashed line indicates level of 95% coverage.

### Interpreting the Nonpareil outputs

A subset from each individual sample was analysed using the nonpareil programme. The algorithm was set to kmr in order to save time. Subsets of 10,000 reads were then compared to measure the redundancy of samples. Based on this, a graph could be created, showing sequencing depth.

On figure 1, only three samples have been included but generally for most of the samples, sequencing was not deep enough as the solid curves do not reach the 0.95 interval. This could be due to insufficient sample sizes or there were too many bacteria in the samples for the amplicon sequencing to work properly. Furthermore, the samples were not denoised which can lead to the false impression that sequencing depth was not sufficient. When sequencing soil, nonpareil suggests that around 200 Gb are necessary to achieve 95% coverage. For faeces, this number is significantly lower. For our samples, the largest sample was around 3 Gb which means that it might not be sufficient to cover the entire metagenome. To achieve higher sequencing depth, larger samples would be ideal. Additionally, several samples should be taken from the same drain to ensure adequate sampling and eliminate errors [3, 4].

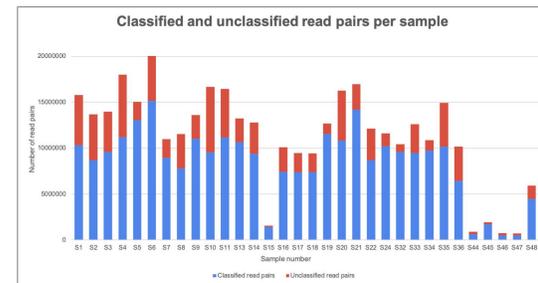


Figure 2: Number of total read pairs for each sample. The blue bars indicate how many of the pairs the Kaiju software was able to classify into a taxonomic unit. Most samples ranged between 10-30% unclassified read pairs.

### Read numbers and Kaiju classification

The read pair of all samples were run through the Kaiju programme, which then attempted to match each pair to a database entry. Figure 2 thus illustrates how well Kaiju managed to classify each sample.

It can be seen that a majority of samples consisted of over 10<sup>6</sup> read pairs. However, 5 of the samples have significantly fewer reads. This could be due to low concentrations of DNA in the samples. As such, these samples likely do not fully represent each of their respective metagenomes. In the case of most samples, a substantial number of read pairs remain unclassified. This is likely due to the presence of a large number of organisms that yet have to be cultured, sequenced or implemented into databases. As such, these unclassified reads might represent a piece of the metagenome which cannot be studied yet.

## Phyla distribution of drain ecosystems

After Kaiju had classified a majority of each sample, it was possible to examine the resulting data to get an idea of the taxonomic distributions in the sampled drain systems. An overview of each sample and their respective drain system of origin can be seen in table 1. In the case of the samples S17, S18, S33, and S35, it was not possible to estimate the age of their sample drain systems. As such, they were omitted from further analysis, so as not to obscure the data.

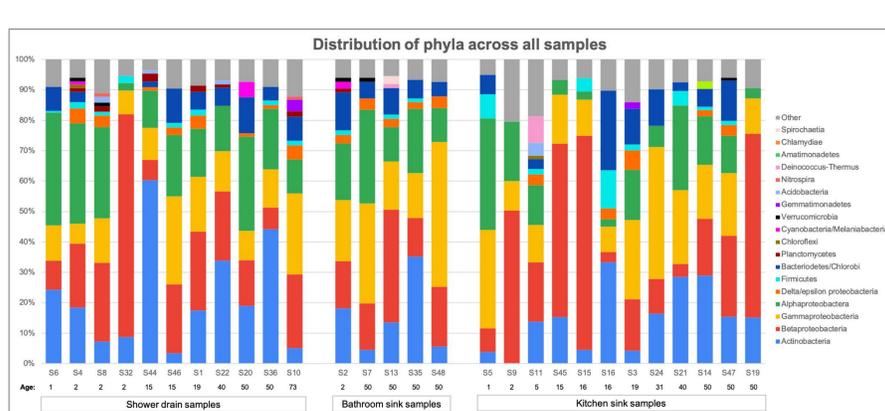
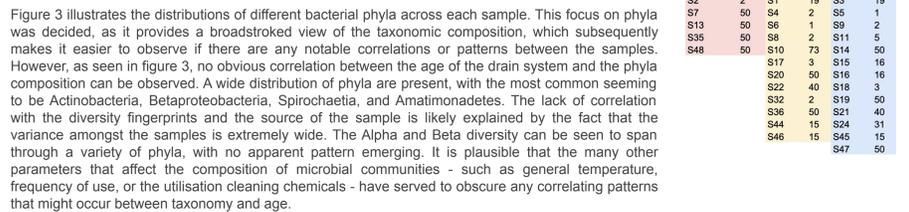


Figure 3: The distribution of phyla across samples, organised with respect to the age and typing of each respective drain system of origin. Phyla that contributed to less than 1% of the total taxonomic composition were merged into the class named 'Other'.

## Principal Component Analysis

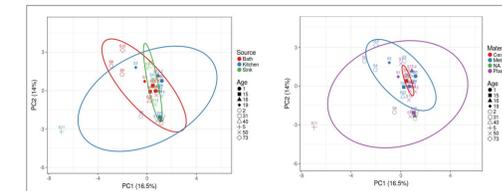


Figure 4: Principal Component Analysis: Unit variance scaling is applied to rows; SVD with imputation is used to calculate principal components. X and Y axis show principal component 1 and principal component 2 that explain 16.5% and 14% of the total variance, respectively. Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse. N = 28 data points. Unit variance scaling uses the square root of the SD for scaling.

While at first glance, there seems to be a statistically significant indication that kitchens and plastic drains can be identified via their location on the PC1 axis, there seems to be a single statistical outlier (S11) producing this result. It should be noted that this drain contains *Methiobacterium silvanus*, a thermophile. This corresponds with the fact that hot water is often used in kitchen sinks.

## Species richness and diversity

Shannon-index: Index to determine the diversity in samples. The lower the index, the lower the diversity. They usually range from 1.5 to 3.5.

To compare the composition of the samples and find out if age makes a difference, we computed the species richness and the diversity (by Shannon-index) and the resulting diagrams are shown to the right. The samples are arranged in order of the youngest to the left and oldest to the right.

First, it should be noted that we have few samples, meaning that even if there appeared to be a difference, it would not be statistically significant. Furthermore, the Shannon indexes are all between 1 and 2 - some exceeding below the normal range of the Shannon index. This means that the diversity in those samples is very low. Half of the samples have a diversity above 1.5 and they are most abundant in isolates from the kitchen. It is quite evident from the Shannon indexes that no correlation can be observed as they all seem random.

Looking at the Kitchen and Bath samples, it seems that the Bath samples generally have a higher richness. The Bathroom sink samples are disregarded because of the low amount.

## Bray-Curtis dissimilarity

Bray-Curtis dissimilarity: Statistic used to quantify the compositional dissimilarity between two different sites, based on counts at each site.

$$C_i = \text{sum of the lowest count of all common species}$$

$$S_i + S_j = \text{total count of the sample}$$

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

To compare the microbial composition of samples from similar environments, we created Bray-Curtis dissimilarity heat-maps for the three environments: Shower, bathroom sink, and kitchen. The heat-map from the bath samples show a low dissimilarity between many of the samples (S20, S2, S22, S1, S6, S36, S10 and S8), whereas S32, S46 and S44 have a high dissimilarity compared to any sample. Samples from the bathroom sink are generally very dissimilar except for between S2 and the samples S35 and S13. The samples from the kitchen generally have higher dissimilarity than the samples from the bathroom-drain, which correlates with the findings from the analysis of phyla richness and Shannon-index from the two environments, as the bath generally had higher phyla-richness whereas the kitchen generally had a higher Shannon-index.

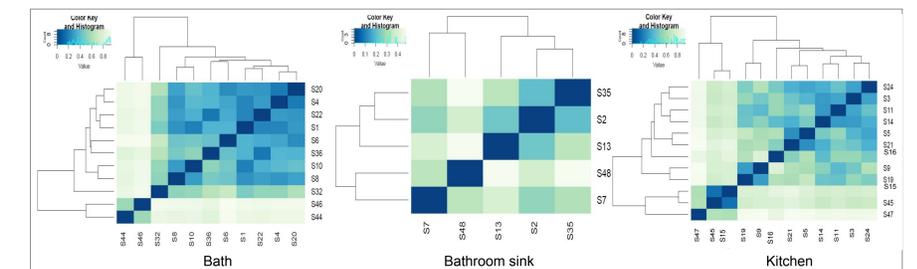


Figure 5: Heat-maps showing the Bray-Curtis dissimilarity of samples from different environments: bath, bathroom sink, and kitchen.

## Discussion and Conclusion

A previous study has demonstrated the presence of especially Proteobacteria and, to a lesser degree, Firmicutes [1]. This was backed up by our findings. Another study had found *Helicobacter* in 91% of samples from cold water taps [2]. Contrastingly, our data showed only very few *Helicobacter* spp. This could perhaps be explained by the temperature of the water which is being poured into the drain on a daily basis: both showers and kitchen sinks generally use warm water. Additionally, this could perhaps be explained by the fact that sampling methods and materials vary.

While this study has examined the diversity of bacterial phyla in the samples, not much can be said about the abundance of each bacterium. For future work, it would be interesting to look not only at metagenomics data but also qPCR of the 16S rRNA sequences to find the number of bacteria.

Additionally, it would be interesting to look at the same drains several times from they are installed until 10-15 years after in order to investigate the development of biofilm over time in the same drain. This would give a better picture of the development of biofilm over time with fewer variables to influence the data.

In conclusion, no significant differences in biofilm composition based on the age of the pipes were observed. However, we saw an indication that the material of the pipes might have an influence on biofilm composition. In order to achieve more certain and reliable results, more experiments are needed.

## References

- [1] McBain, Andrew J., Robert G. Bartolo, Carl E. Catrenich, Duane Charbonneau, Ruth G. Ledder, Alexander H. Rickard, Sharon A. Symmons, and Peter Gilbert. 2003. "Microbial Characterization of Biofilms in Domestic Drains and the Establishment of Stable Biofilm Microcosms." Applied and Environmental Microbiology 69 (1): 177-85.
- [2] Elbigodin, Kevin E., Alyson Smith, and Catherine Anne Biggs. 2006. "A Review of Biofilms in Domestic Plumbing." Journal of American Water Works Association 100 (10): 131-38+12, 12-138.
- [3] Rodriguez-R et al. 2018. Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. mSystems 3(3): e00039-18. DOI: 10.1128/mSystems.00039-18.
- [4] Rodriguez-R & Konstantinidis. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30 (5): 629-635. DOI: 10.1093/bioinformatics/btt584.
- [5] <https://bit.cs.ut.ee/custjv/>