University of Applied Sciences and Arts Northwestern Switzerland
School of Life Sciences

Bacterial Diversity In Five Different Waste Water Treatment Plants

Begrich, Annette

Bachelor thesis, Medical Computer Sciences

Principal: Prof. Dr. G. Lipps, ICB FHNW
Expert: Dr. B. Kolvenbach, IEC FHNW

ABSTRACT

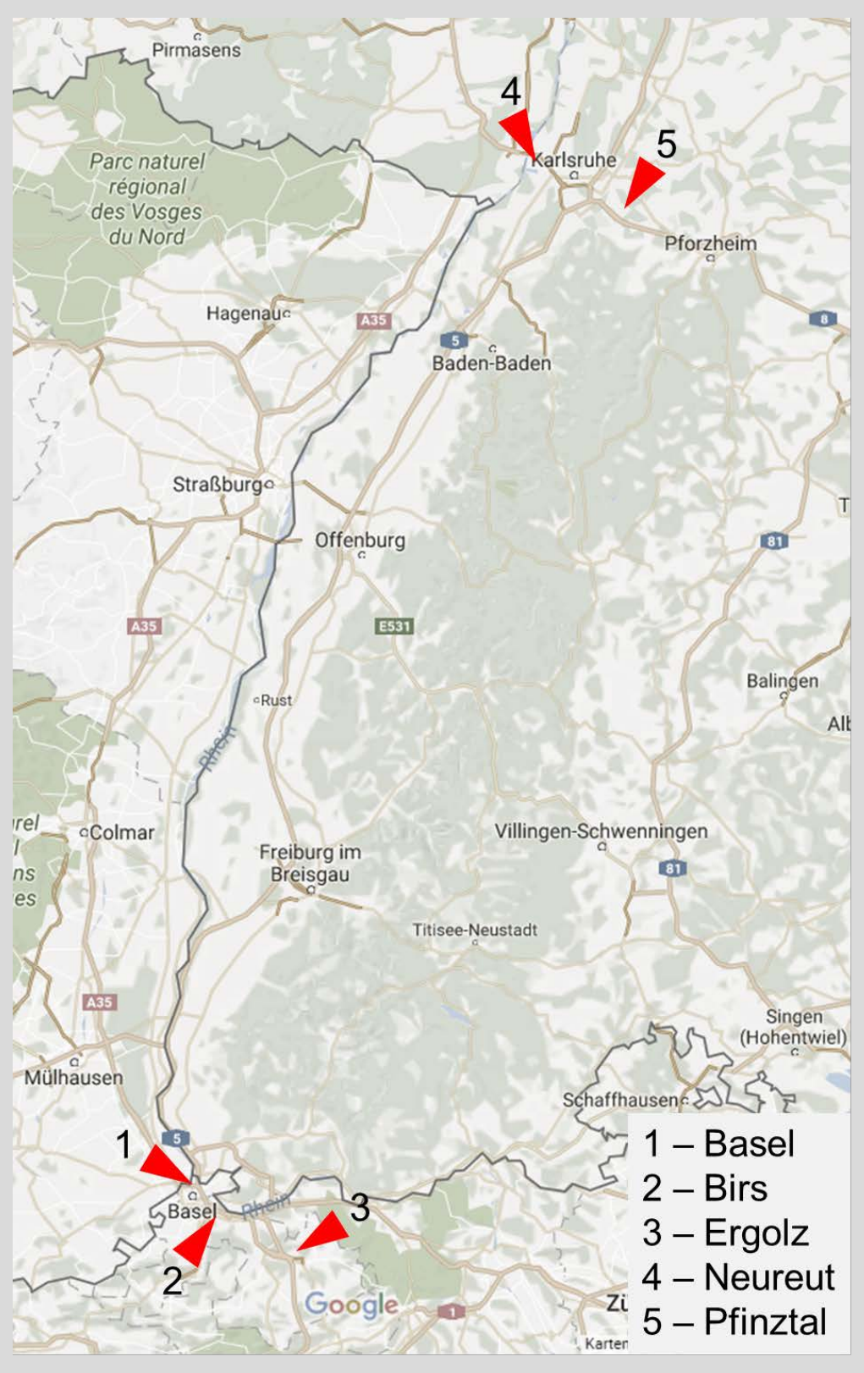


Fig. 1 Locations

Biological Waste Water (WW) and its composition has been of interest to ecological scientists in recent decades on different levels. The analysis of the bacterial composition has been primarily dependent on their culturability and therefore it was strongly biased towards this aspect. Recent advances, especially in molecular biology technologies, have provided further understanding into the complex interactions.

To analyze the bacterial communities of five different waste water treatment plants (WWT) in southern Germany and northwestern Switzerland sludge was collected over the span of a year, the hypervariable regions V3 and V4 of the 16S gene were amplified and sequenced. The 16S gene is used to deconvolute bacterial communities due to the high preservation of the nine hypervariable regions, which can be used to identify the bacteria and their phylogeny. In the scope of this bachelor thesis, sequencing data from these WWT was processed through two different data processing pipelines and compared to each other. In addition, generated data was analyzed for bacterial composition and differences based on seasonal and geographical differences .

INTRODUCTION

Earlier methods to analyze the bacterial composition were dependent on the ability of bacteria to be cultured and hence were strongly biased towards that specific aspect. Only recently, more suitable analysis methods have been developed giving us a better representation of the different bacteria present in WW. In particular, molecular biological techniques have provided further insights into the complex interaction and bacterial composition of waste water. However, low abundant bacterial strains were rarely identified and often lost in the 'noise'.

With the development of the Sanger chain termination sequencing method in 1977, the ability to sequence DNA more reliably and reproducibly has helped science to understand many previously unknown mechanisms. Parallel to advances in sequencing, computing power available to researchers also grew exponentially giving us the ability to process information stored in DNA faster and more precisely (1). A new approach to sequencing, termed "next generation sequencing" (NGS), has revolutionized genomic research; thousands of DNA fragments, short reads, are being sequenced simultaneously advancing the rapid identification of highly complex bacterial communities (2).

For the massive amount of data generated in NGS, many different tools exist to quantify and compare the data. Until recently, the first step in analyzing the data was to cluster sequences and assign operational taxonomic units (OTU) from a reference data base (3; 4).

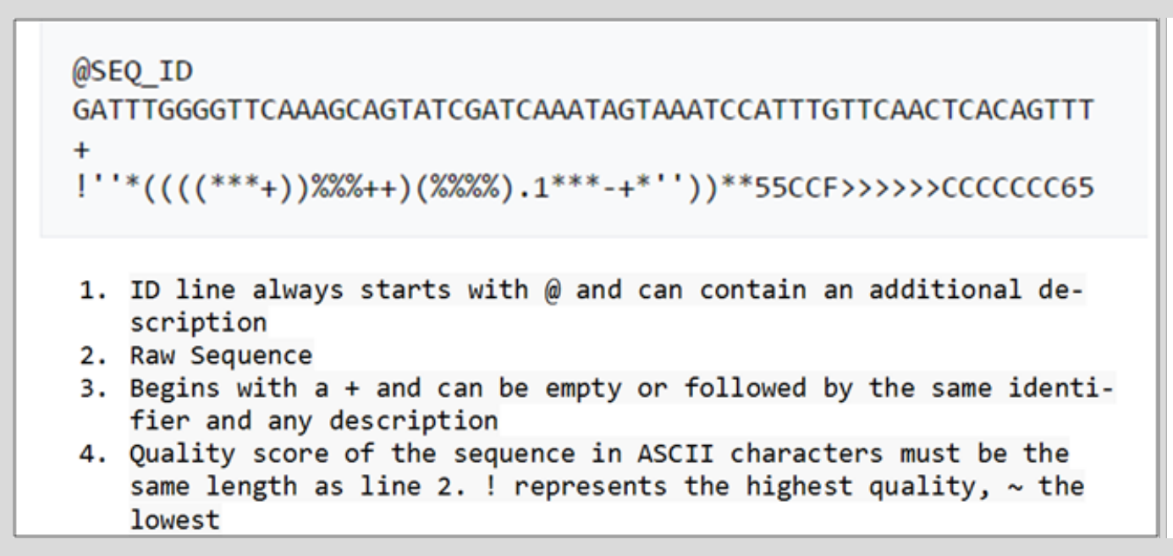
In 2012 Susan P Holmes's group at Stanford University started to develop the Divisive Amplicon Denoising Algorithm (DADA), and in 2016 published the DADA2 R package (5; 6). This new approach takes Illumina amplicon errors into consideration and implements first a filtering, then a dereplication followed by sample interference, before merging paired end reads and assigning a taxonomic unit. This process makes it much more precise for the data analysis. This more detailed approach allows for more precise variation of population structure and can help to identify more complex microbiomes (5).

RESULTS

SEQUENCE QUALITY COMPARISON

One of the first steps in the DADA2 pipeline is to plot the quality of the sequences (3). This information is encoded in the .fastq files from the run

The higher and more consistent the quality of the sequence reads is over the whole length of the sequence the better it can be paired, processed, and used to build the phylogenic tree. The quantity of reads per sample was high, with an average of 301415(±112310), the quality is quite low. This was particularly evident for the reverse sequences. It was therefore decided to be less restrictive with the cut off. A quality score of 25 was selected to approximate the nucleotide pair where the mean (green solid line (3)) dropped below the limit to select the maximum length of the sequences analyzed. This lower threshold was also required to achieve an overlap of the sequences, since the final PCR product was roughly 460bp long. The filter was set to 270 for the forward reads and 230 for the reverse reads



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. ID line always starts with @ and can contain an additional description
2. Raw Sequence
3. Begins with a + and can be empty or followed by the same identifier and any description
4. Quality score of the sequence in ASCII characters must be the same length as line 2. ! represents the highest quality, ~ the lowest

Fig. 2 fastQ format

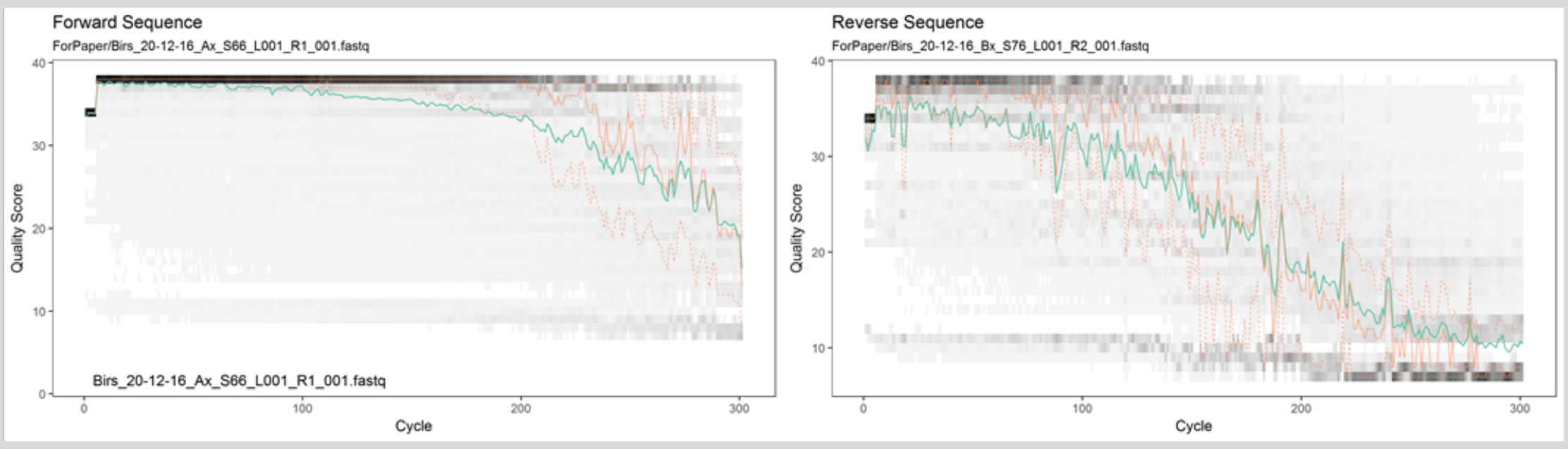
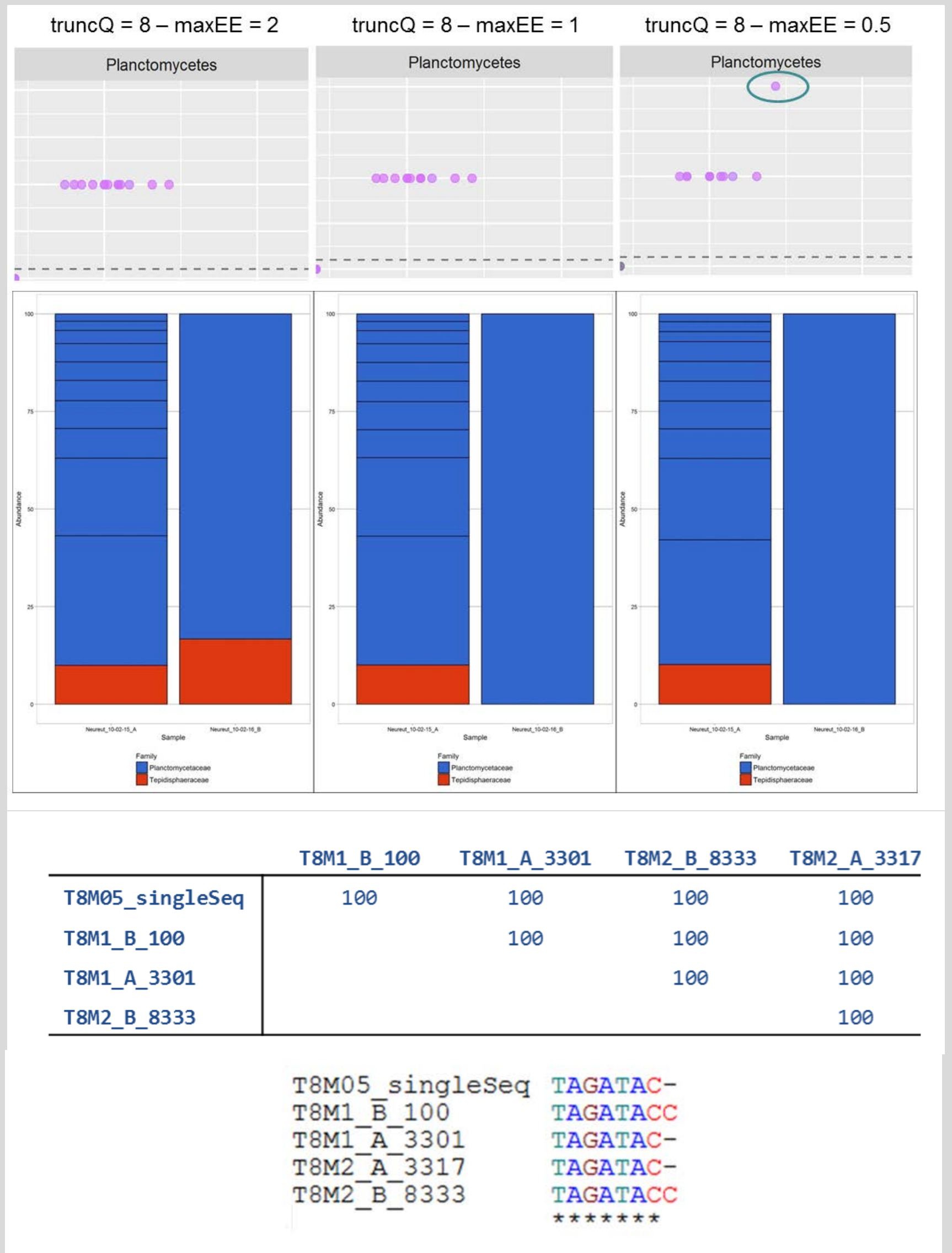


Fig.3 Forward and reverse quality



	T8M1_B_100	T8M1_A_3301	T8M2_B_8333	T8M2_A_3317
T8M05_singleSeq	100	100	100	100
T8M1_B_100		100	100	100
T8M1_A_3301			100	100
T8M2_B_8333				100

```
T8M05_singleSeq TAGATAC-
T8M1_B_100      TAGATACC
T8M1_A_3301     TAGATAC-
T8M2_A_3317     TAGATAC-
T8M2_B_8333     TAGATACC
                *******
```

Fig. 4 Comparison of different filters

COMPARISON OF TECHNICAL REPEATS AND DIFFERENT FILTERS

Due to the lower quality the samples were analyzed with different filtering parameters to see how much this changes or improves the ability to interpret our data. First the top 250 OTU's of each dataset were plotted to see if now more OTU's show a prevalence of 1. As a second step the data was subset to the Planctomycetes and the most abundant sequences were aligned to see if we were able to see if a taxa could be exactly the same in the corresponding repeat.

With the parameters truncQ set at 8 and maxEE set at 0.5 one single sequence is found to be identical. In addition this sequence is the one found in all filtered sets with the highest score for Planctomycetes, just that this time the single nucleotide, in other settings extending on the 3' end, is cut in Repeat B.

DIVERSITY IN LOCATION AND SEASON

The overall diversity of the samples is a good place to start for the analysis. This already shows us that there is little variance between the different locations. While the more industrial, urban WWT's (e.g. Basel and Neureut) had a slightly lower diversity than the more rural plants (e.g. Birs, Ergolz, Pfinztal), this difference was not statistically significant. At this point it is not clear why this minor difference can be observed and there could be many different reasons influencing the diversity.

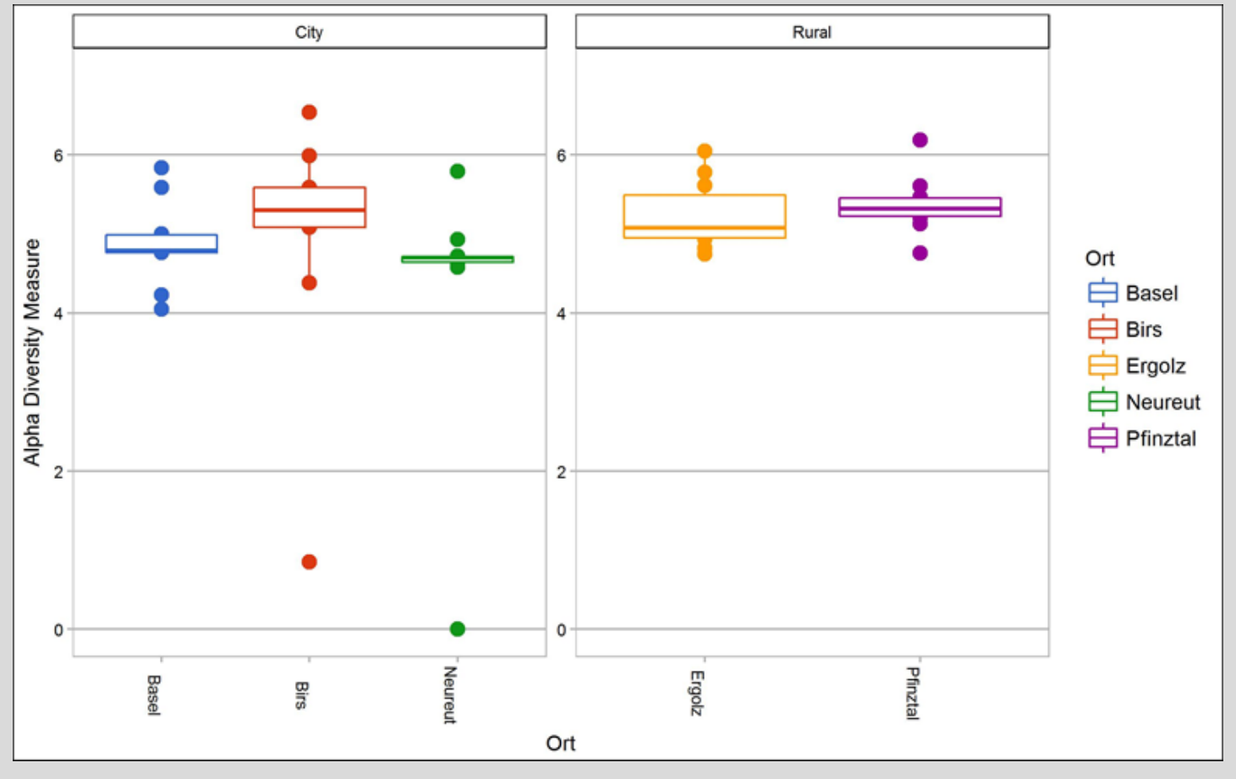


Fig. 5 Diversity by location

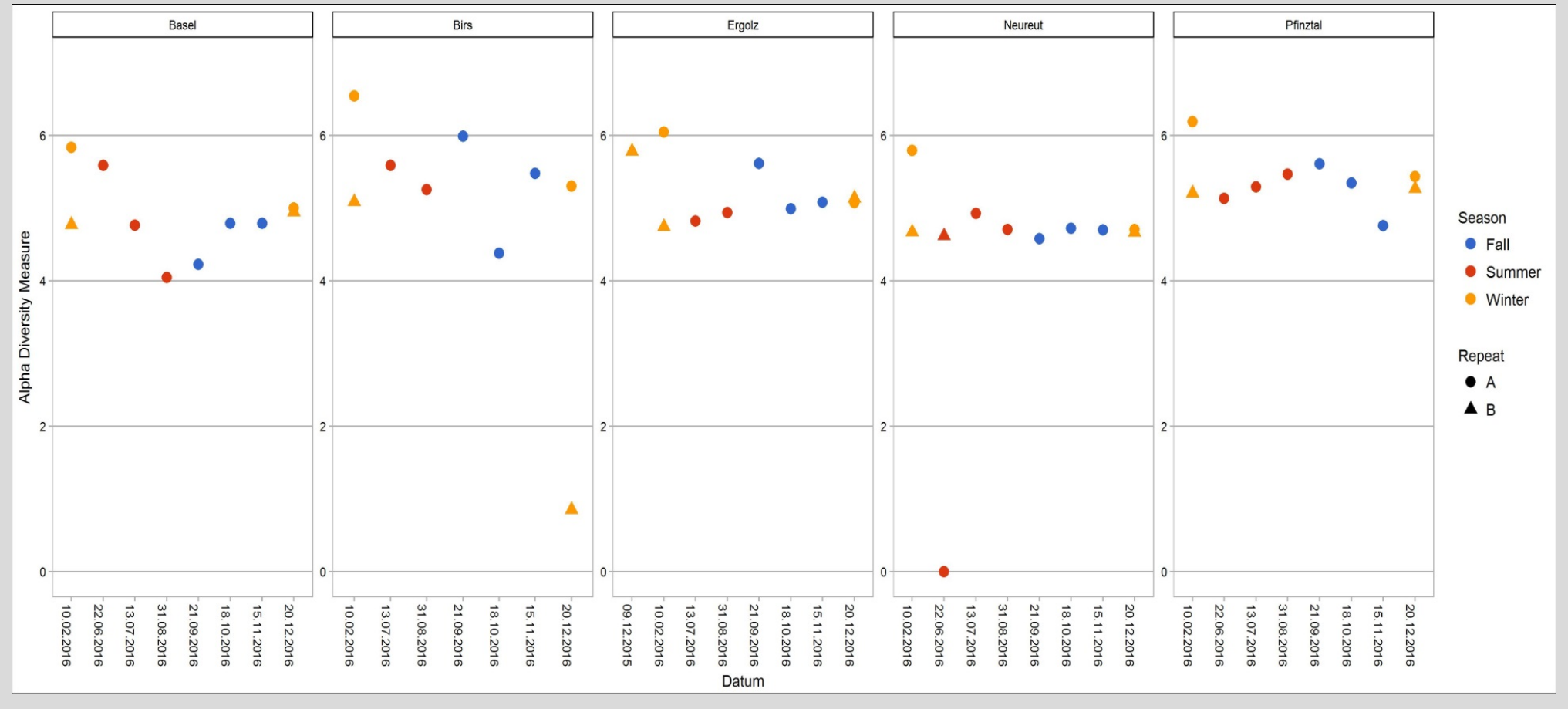


Fig. 6 Diversity by season

Displaying the alpha diversity by plant and date shows us that the diversity in Basel is lower at the end of the summer and higher variety in in the winter months. In contrast the bacterial diversity in Neureut remains quite stable over the whole period investigated while Pfinztal behaves contrary to Basel showing a greater diversity in the summer months. The overall diversity in the Ergolz and Birs plants is very variable and no clear trend could be observed.

The interpretation of seasonal fluctuations in bacterial diversity would be more precise and conclusive, if more samples were taken throughout the year, or if data from more than one year was available. In summary, although one can state that the waste water in Basel had a lower bacterial diversity on the 31st of August and 29th of September in 2016, it can't be clearly stated that the community has a lower diversity at the end of summer in general. It should be noted that any conclusions drawn must be taken with caution since the quality of the reads was lower than expected and it is not clear how the filtering influences the overall diversity

CONCLUSION

The work on a dataset like the one presented in this thesis is never really complete. One can always discover new interpretations that have previously been overlooked. There are also endless ways to graphically visualize the data; the ones used in this project are only a selection of what is possible. The multitude of data collected in one single run of the Illumina MiSeq is astonishing, however, it is important to produce high quality data especially when such a high output can be generated in one run. The quality produced in the scope of this project is suboptimal for a clear interpretation, nevertheless it can be used as a starting point for discussion.

An understanding of seasonal change on the bacterial community of the waste water can be looked at but significant interpretation at this stage is tricky. A more reliable explanation could be achieved if data from a wider timespan were available. This might help to elucidate whether certain changes occur in regular patterns or are just a single occurrence due to specific environmental impacts; these external factors have been disregarded for this dataset, or are simply not known. The data produced and the processing pipeline investigated in this project are a beginning and more could be done time and resources permitting.

REFERENCES

1. *An Introduction to Next-Generation Sequencing Technology.* 2016, **Illumina.**.

2. *Illumina-based analysis of mircobial community diversity.* **Degnan, PH; Ochman, H.** 2012, The ISME Journal.

3. *Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses.* **Callahan, Ben J, et al., et al.** 2016, F1000Research.

4. *A perspective on 16S rRNA operational toxonomic unit clustering using sequence similarity.* **Nguyen, Nam-Phuong, et al., et al.** 2016, npj Biofilms and Micorbiome.

5. *DADA2: High-resolution sample inference form Illumina amplicon data.* **Callahan, Ben J et al.** 2015, Nature Methods

6. *Denoising PCR-amplified metagenome data.* **Rosen, M, et al., et al.** 2012, BMC Bioinformatics.