Supplementary Material

Small-scale heterogeneity in drinking water biofilms

Lisa Neu\(^1,2\), Caitlin R. Proctor\(^1,3\), Jean-Claude Walser\(^4\), Frederik Hammes\(^1\*)

\(^1\) Department of Environmental Microbiology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
\(^2\) Department of Environmental Systems Science, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Zürich, Switzerland
\(^3\) Schools of Civil, Environmental and Ecological, Materials, and Biomedical Engineering, Purdue University, West Lafayette, Indiana, United States
\(^4\) Genetic Diversity Centre (GDC), ETH Zürich, Zürich, Switzerland

Table of contents

- **Figure S1** Temperature pattern in control hose during operation
- **Figure S2** Thickness values for control hose biofilm on 2mm-scale
- **Figure S3** Scanning electron microscopy images (control & real Hose)
- **Figure S4** NMDS plots for comparisons on DNA-level
- **Figure S5** Correlations between three dominant taxa (control hose)
- **Figure S6** Thickness values for real hose biofilm on 2mm-scale
- **Figure S7** Correlations between three dominant taxa (real hose)
- **Figure S8** TCC for different, theoretically possible sampling sizes
- **Table S1** Water characteristics
- **Table S2** Settings for Amplification PCR and INDEX PCR reactions
- **Table S3** Processing of 16S rRNA gene sequences
- **Table S4** Dominant taxa in control hose biofilm
- **Table S5** Dominant taxa in real hose biofilm
Supplementary Figure S1. Temperature pattern inside PVC-P hose (black) and surrounding temperature inside box (gray). (A) over the course of 1 week; in higher resolution, the (B) 9 am flush and (C) 5 pm flush.
Supplementary Figure S2. Thickness of the control hose biofilm. Optical coherence tomography was used for imaging and analyzing structure and thickness of a biofilm grown inside a flexible PVC-P hose under controlled conditions in the laboratory. Images were taken two-dimensional in 2 mm length and 1 mm in height. Here, each bar represents the average thickness for these 2 mm-sections for 1.20 m hose length.
Supplementary Figure S3. Scanning electron microscopic image of biofilms grown under (A) controlled laboratory conditions (Control hose) or (B) real use (i.e., uncontrolled) conditions (Real hose). Images made by the Center for Microscopy and Image Analysis, University of Zurich.
**Supplementary Figure S4.** Non-metric multidimensional scaling representation of bray-curtis dissimilarity between biofilm communities, either grown under controlled laboratory conditions (control hose, A) or under real (i.e., uncontrolled) use conditions (real hose, B). NMDS plots display dissimilarity between subsamples of the control hose biofilm (A), between subsamples of the real hose biofilm (B), and between both biofilms (C).
**Supplementary Figure S5.** Correlations between the relative abundances of the three most dominant taxa in the control hose biofilm. (A) top, (B) bottom part of the hose. Dominant taxa were identified as *Bradyrhizobium* spp., Cytophagaceae, and TM6_(Dependentiae).
Supplementary Figure S6. Thickness of the real hose biofilm. Optical coherence tomography was used for imaging and analyzing structure and thickness of a biofilm grown inside a flexible PVC-P hose under real (i.e., uncontrolled) use conditions. Images were taken two-dimensional in 2 mm length and 1 mm in height. Here, each bar represents the average thickness for these 2 mm-sections for 1.20 m hose length.
Supplementary Figure S7. Correlations between the relative abundances of the three most dominant taxa in the real hose biofilm. (A) top, (B) bottom part of the hose. Dominant taxa were identified as *Bradyrhizobium* spp., *Altererythrobacter* spp., and *Caulobacter* spp.
Supplementary Figure S8. Relevance of sample size. Bacterial cell numbers of the control hose biofilm were exemplarily used to assess the importance of sample size for the representativeness of results. Samples have been created on 1.2 cm-scale. Here, several sections have been combined to illustrated larger sampling sizes. Black dots represent the overall average for the entire shower hose biofilm. Grey dots illustrate the variation of results, decreasing with increasing sample size (i.e., samples cm in length).
Supplementary Table S1. Water characteristics for control and real hose biofilm.

<table>
<thead>
<tr>
<th>(A) Control hose</th>
<th>(B) Real hose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Origin</strong></td>
<td></td>
</tr>
<tr>
<td>15.0 %</td>
<td>96%</td>
</tr>
<tr>
<td>Ground water</td>
<td>6.4 %</td>
</tr>
<tr>
<td>Spring water</td>
<td>4%</td>
</tr>
<tr>
<td>Lake water</td>
<td>-</td>
</tr>
<tr>
<td><strong>Physico-chemical parameters</strong></td>
<td></td>
</tr>
<tr>
<td>95 mg/L Calcium</td>
<td>100 mg/L</td>
</tr>
<tr>
<td>12 mg/L Natrium</td>
<td>5 mg/L</td>
</tr>
<tr>
<td>14.7 mg/L Magnesium</td>
<td>15 mg/L</td>
</tr>
<tr>
<td>14.3 mg/L Nitrate</td>
<td>21 mg/L</td>
</tr>
<tr>
<td>1.9 mg/L Potassium</td>
<td>-</td>
</tr>
<tr>
<td>15.6 mg/L Chloride</td>
<td>14 mg/L</td>
</tr>
<tr>
<td>16.7 mg/L Sulfate</td>
<td>20 mg/L</td>
</tr>
<tr>
<td>0.05 mg/L Fluoride</td>
<td>&lt; 0.1 mg/L</td>
</tr>
<tr>
<td>- Manganese</td>
<td>&gt; 0.01 mg/L</td>
</tr>
<tr>
<td>- Hydrogen carbonate</td>
<td>31 mg/L</td>
</tr>
<tr>
<td>- Iron</td>
<td>&lt; 0.01 mg/L</td>
</tr>
<tr>
<td>- pH</td>
<td>7.3</td>
</tr>
<tr>
<td>26.8 °fH Hardness</td>
<td>34 °fH</td>
</tr>
<tr>
<td>13.2 °C Temperature</td>
<td>12 °C</td>
</tr>
</tbody>
</table>
Supplementary Table S2. Detailed information on PCR reactions

<table>
<thead>
<tr>
<th>A – Amplicon PCR</th>
<th>Volume (25 µL reaction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x KAPA HiFi HotStart ReadyMix</td>
<td>12.5 µL</td>
</tr>
<tr>
<td>Forward primer (10 µM)</td>
<td>0.75 µL</td>
</tr>
<tr>
<td>Reverse primer (10 µM)</td>
<td>0.75 µL</td>
</tr>
<tr>
<td>Template DNA (adjusted to 1 ng with Dnase free water)</td>
<td>11.0 µL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Duration</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95 °C</td>
<td>5:00 min</td>
<td></td>
</tr>
<tr>
<td>95 °C</td>
<td>0:20 min</td>
<td></td>
</tr>
<tr>
<td>51 °C</td>
<td>0:15 min</td>
<td>29 x</td>
</tr>
<tr>
<td>72 °C</td>
<td>0:30 min</td>
<td></td>
</tr>
<tr>
<td>4 °C</td>
<td>hold</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B – Index PCR</th>
<th>Volume (50 µL reaction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x KAPA HiFi HotStart ReadyMix</td>
<td>25.0 µL</td>
</tr>
<tr>
<td>Nextera XT Index 1 primer</td>
<td>5.0 µL</td>
</tr>
<tr>
<td>Nextera XT Index 2 primer</td>
<td>5.0 µL</td>
</tr>
<tr>
<td>Template DNA</td>
<td>15.0 µL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Duration</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>95 °C</td>
<td>3:00 min</td>
<td></td>
</tr>
<tr>
<td>95 °C</td>
<td>0:30 min</td>
<td></td>
</tr>
<tr>
<td>51 °C</td>
<td>0:35 min</td>
<td>8 x</td>
</tr>
<tr>
<td>72 °C</td>
<td>0:35 min</td>
<td></td>
</tr>
<tr>
<td>4 °C</td>
<td>hold</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table S3. Processing of 16S rRNA gene sequences.

<table>
<thead>
<tr>
<th>(A) Quality Control</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FastQC V0.11.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B) Trimming and merging of primer sites</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>usearch v10.0.240_i86linux64</td>
<td></td>
</tr>
<tr>
<td>Trim R1 20</td>
<td></td>
</tr>
<tr>
<td>Trim R2 50</td>
<td></td>
</tr>
<tr>
<td>Flash v1.2.11</td>
<td></td>
</tr>
<tr>
<td>Minimal overlap 15</td>
<td></td>
</tr>
<tr>
<td>Maximal overlap 300</td>
<td></td>
</tr>
<tr>
<td>Maximal mismatch density 0.25</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(C) Primer Site Trimming</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Usearch v10.0.240_i86linux64</td>
<td></td>
</tr>
<tr>
<td>Coverage full-length</td>
<td></td>
</tr>
<tr>
<td>Allowed number of mismatches 1</td>
<td></td>
</tr>
<tr>
<td>Amplicon size range 50 - 600</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(D) Filtering based on quality and size</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size range 200 - 500</td>
<td></td>
</tr>
<tr>
<td>GC range 30 - 70</td>
<td></td>
</tr>
<tr>
<td>Minimal Q mean 200 - 500</td>
<td></td>
</tr>
<tr>
<td>Number of Ns 1</td>
<td></td>
</tr>
<tr>
<td>Low complexity dust / 30</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table S4. List of all dominant taxa (i.e., with at least 1 % of the total number of reads) in the control hose biofilm.

<table>
<thead>
<tr>
<th>Portion of total community (%)</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.7</td>
<td>Bacteroidetes</td>
<td>Cytophagia</td>
<td>Cytophagales</td>
<td>Cytophagaceae</td>
<td>uncultured</td>
</tr>
<tr>
<td>23.4</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhizobiales</td>
<td>Bradyrhizobiaceae</td>
<td>Bradyrhizobium</td>
</tr>
<tr>
<td>9.6</td>
<td>TM6_[Dependentiae]</td>
<td>uncultured_bacterium</td>
<td>uncultured_bacterium</td>
<td>uncultured_bacterium</td>
<td>uncultured_bacterium</td>
</tr>
<tr>
<td>8.6</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhodobacterales</td>
<td>Rhodobacteraceae</td>
<td>NA</td>
</tr>
<tr>
<td>6.3</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhodobacterales</td>
<td>Rhodobacteraceae</td>
<td>Rhodobacter</td>
</tr>
<tr>
<td>6.2</td>
<td>Bacteroidetes</td>
<td>Sphingobacteria</td>
<td>Sphingobacterales</td>
<td>Chitinophagaceae</td>
<td>Sediminibacterium</td>
</tr>
<tr>
<td>5.7</td>
<td>Verrucomicrobia</td>
<td>Verrucomicrobia</td>
<td>Verrucomicrobiales</td>
<td>Verrucomicrobiaceae</td>
<td>Brevifollis</td>
</tr>
<tr>
<td>1.9</td>
<td>Bacteroidetes</td>
<td>Cytophagia</td>
<td>Cytophagales</td>
<td>Cytophagaceae</td>
<td>Ohtaekwangia</td>
</tr>
<tr>
<td>1.5</td>
<td>Proteobacteria</td>
<td>Betaproteobacteria</td>
<td>Rhodocyclales</td>
<td>Rhodocyclaceae</td>
<td>Dechloromonas</td>
</tr>
<tr>
<td>1.4</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Caulobacterales</td>
<td>Caulobacteraceae</td>
<td>Phenyllobacterium</td>
</tr>
<tr>
<td>1.3</td>
<td>Proteobacteria</td>
<td>Betaproteobacteria</td>
<td>Rhodocyclales</td>
<td>Rhodocyclaceae</td>
<td>Denitratisoma</td>
</tr>
</tbody>
</table>
**Supplementary Table S5.** List of all dominant taxa (i.e., with at least 1% of the total number of reads) in the real hose biofilm.

<table>
<thead>
<tr>
<th>Portion of total community (%)</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.72</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Caulobacterales</td>
<td>Caulobacteraceae</td>
<td>Caulobacter</td>
</tr>
<tr>
<td>24.24</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhizobiales</td>
<td>Bradyrhizobiaceae</td>
<td>Bradyrhizobium</td>
</tr>
<tr>
<td>14.24</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Sphingomonadales</td>
<td>Erythrobacteraceae</td>
<td>Altererythrobacter</td>
</tr>
<tr>
<td>5.41</td>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Micrococcales</td>
<td>Brevibacteriaceae</td>
<td>Brevibacterium</td>
</tr>
<tr>
<td>4.79</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhizobiales</td>
<td>Bradyrhizobiaceae</td>
<td>Bosea</td>
</tr>
<tr>
<td>1.95</td>
<td>Bacteroidetes</td>
<td>Sphingobacteriia</td>
<td>Sphingobacteriales</td>
<td>Chitinophagaceae</td>
<td>NA</td>
</tr>
<tr>
<td>1.36</td>
<td>Proteobacteria</td>
<td>Deltaproteobacteria</td>
<td>Bdellovibrionales</td>
<td>Bdellovibrionaceae</td>
<td>Bdellovibrio</td>
</tr>
<tr>
<td>1.35</td>
<td>Bacteroidetes</td>
<td>Sphingobacteriia</td>
<td>Sphingobacteriales</td>
<td>Chitinophagaceae</td>
<td>uncultured</td>
</tr>
<tr>
<td>1.16</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Sphingomonadales</td>
<td>Sphingomonadaceae</td>
<td>Sphingomonas</td>
</tr>
<tr>
<td>1.15</td>
<td>Cyanobacteria</td>
<td>ML635J-21</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1.14</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhodobacterales</td>
<td>Rhodobacteraceae</td>
<td>Rhodobacter</td>
</tr>
</tbody>
</table>
Biofilm heterogeneity has been characterized on various scales for both natural and engineered ecosystems. This heterogeneity has been attributed to spatial differences in environmental factors. Understanding their impact on localized biofilm heterogeneity in building plumbing systems is important for both management and representative sampling strategies. We assessed heterogeneity within the confined engineered ecosystem of a shower hose by high-resolution sampling (200 individual biofilm sections per hose) on varying scales (µm to m), postulating that a biofilm grown on a single material under uniform conditions should be homogeneous in its structure, bacterial numbers, and community composition. A biofilm grown for 12 months under controlled laboratory conditions, showed homogeneity on large-scale. However, small-scale heterogeneity was clearly observed. For example, biofilm thickness of cm-sections varied up to 4-fold (150 – 750 µm), total cell concentrations (TCC) 3-fold (1.1 – 3.4 x 10^7 cells/cm2), and dominant taxa showed fluctuating trends and localized heterogeneity. A biofilm grown under real (i.e., uncontrolled) conditions developed clearly more heterogeneity on both large- and small-scale, with, e.g., impressive spatial fluctuations of the most dominant taxa. Interestingly, both biofilm communities showed comparably low diversities, with < 400 taxa each, and solely 3 taxa accounting for 57 % and 73 % of the respective community. This low diversity was attributed to a strong selective pressure, originating in migrating carbon from the flexible hoses as major carbon source. High-resolution sampling on various scales allowed for a detailed analysis of spatial heterogeneity in drinking water biofilms. This study gives insight into biofilm structure and community composition on cm-to m-scale and is particularly useful for decisions on sampling strategies in biofilm research and monitoring.

Set R Environment

```r
## RESET R ENVIRONMENT
rm(list=ls()) # clean/reset environment

## SET WORKING DIRECTORY
setwd("myfolder")
wd <- getwd()
wd
list.files(wd)
```

Libraries and Packages

```r
require("phyloseq")
require("ggplot2")
require("vegan")
require("microbiome")
```
Data is organised in different categories (highlighted in bold font):

The **Experiment** is subdivided in **Control hose** and **Real hose**, both separated according to the **Orientation** of the hoses (Bottom, Top, "Bottom", "Top"). **Kind** corresponds to samples that were amplified in triplicates for amplification/sequencing control, and **Character** serves for distinguishing between **Experimetal samples** and **Control samples** (e.g., Negative Control for PCR amplification).

```r
# LOAD AND IMPORT DATA
treefile <- "p399_Run171106_Run180108_16S_ZOTU_CLU.tre"
otufile <- "p399_Run171106_Run180108_16S_ZOTU_c97_Count_Sintax.txt"
refseqfile <- "p399_Run171106_Run180108_16S_ZOTU_c97.fa"
mapfile <- "p399_run171106_run180108_MapFile1.txt"

d <- import_qiime(otufilename = otufile, mapfilename = mapfile, treefilename = treefile, refseqfilename = refseqfile)

## DATA

d

# phyloseq-class experiment-level object
# otu_table() OTU Table: [ 545 taxa and 410 samples ]
# sample_data() Sample Data: [ 410 samples by 11 sample variables ]
# tax_table() Taxonomy Table: [ 545 taxa by 7 taxonomic ranks ]
# phy_tree() Phylogenetic Tree: [ 545 tips and 544 internal nodes ]
# refseq() DNAStringSet: [ 545 reference sequences ]

## SUBSET DATA (for "Experiment"; i.e., control and real hose data still combined)
d.e <- subset_samples(d, Character == "e")

d.e

# phyloseq-class experiment-level object
# otu_table() OTU Table: [ 545 taxa and 399 samples ]
# sample_data() Sample Data: [ 399 samples by 11 sample variables ]
# tax_table() Taxonomy Table: [ 545 taxa by 7 taxonomic ranks ]
# phy_tree() Phylogenetic Tree: [ 545 tips and 544 internal nodes ]
# refseq() DNAStringSet: [ 545 reference sequences ]
```

### Explore data
Processing experimental data for analysis

Remove ZOTUs with zero counts and singletons

Samples with low counts might not represent diversity very well and should therefore be removed.

```r
# DETERMINE NUMBER OF ZOTUs WITH <1 AND/OR =1 READS
sum(taxa_sums(d.e) < 1)
# [1] 14

sum(taxa_sums(d.e) ==1)
# [1] 2

# REMOVE ZOTUs WITH <1 AND/OR =1 READS
d.e.noZ <- prune_taxa(taxa_sums(d.e) > 1, d.e)
d.e.noZ
```

**OTU Table:** [529 taxa and 399 samples]
**Sample Data:** [399 samples by 11 sample variables]
**Taxonomy Table:** [529 taxa by 7 taxonomic ranks]
**Phylogenetic Tree:** [529 tips and 528 internal nodes]
**DNAStringSet:** [529 reference sequences]

Set minimum depth and seed
```r
# CHECK MINIMAL DEPTH
min_depth <- min(sample_sums(d.e))
min_depth
# [1] 5648

# CHANGE MINIMAL DEPTH (based on read-count distribution)
min_depth <- c(25000)
min_depth
# [1] 25000

# SET SEED AND RAREFY EVEN DEPTH
set.seed(1)
d.e.rare <- rarefy_even_depth(d.e.noZ, sample.size = min_depth, replace = TRUE, rngseed = 1)

# `set.seed(1)` was used to initialise repeatable random subsampling.
# Please record this for your records so others can reproduce.
# Try `set.seed(1); .Random.seed` for the full vector
# ...
# 10 samples removed because they contained fewer reads than `sample.size`.
# Up to first five removed samples are:
# B005-2B005-3F054F059F080
# ...
# 2OTUs were removed because they are no longer
# present in any sample after random subsampling
# ...

d.e.rare
# phyloseq-class experiment-level object
# otu_table() OTU Table: [ 527 taxa and 389 samples ]
# sample_data() Sample Data: [ 389 samples by 11 sample variables ]
# tax_table() Taxonomy Table: [ 527 taxa by 7 taxonomic ranks ]
# phy_tree() Phylogenetic Tree: [ 527 tips and 526 internal nodes ]
# refseq() DNAStringSet: [ 527 reference sequences ]
```

Subset data in Control and Real hose
# CONTROL HOSE BIOFILM

d.e.control <- subset_samples(d.e.rare, Experiment == "c")
d.e.control

# phyloseq-class experiment-level object
# `otu_table()` OTU Table: [ 527 taxa and 197 samples ]
# `sample_data()` Sample Data: [ 197 samples by 11 sample variables ]
# `tax_table()` Taxonomy Table: [ 527 taxa by 7 taxonomic ranks ]
# `phy_tree()` Phylogenetic Tree: [ 527 tips and 526 internal nodes ]
# `refseq()` DNAStringSet: [ 527 reference sequences ]

# REAL HOSE BIOFILM

d.e.real <- subset_samples(d.e.rare, Experiment == "r")
d.e.real

# phyloseq-class experiment-level object
# `otu_table()` OTU Table: [ 527 taxa and 192 samples ]
# `sample_data()` Sample Data: [ 192 samples by 11 sample variables ]
# `tax_table()` Taxonomy Table: [ 527 taxa by 7 taxonomic ranks ]
# `phy_tree()` Phylogenetic Tree: [ 527 tips and 526 internal nodes ]
# `refseq()` DNAStringSet: [ 527 reference sequences ]

## Export for final analysis of sequencing data outside the R environment

### # EXPORT FOR CONTROL HOSE BIOFILM

ot1 <- as(otu_table(d.e.control), 'matrix')
tax1 <- as(tax_table(d.e.control), 'matrix')
write.csv(ot1, file = "Lineage_Control_1.csv")
write.csv(tax1, file = "Lineage_Control_2.csv")

### # EXPORT FOR REAL HOSE BIOFILM

ot2 <- as(otu_table(d.e.real), 'matrix')
tax2 <- as(tax_table(d.e.real), 'matrix')
write.csv(ot2, file = "Lineage_Reall_1.csv")
write.csv(tax2, file = "Lineage_Real_2.csv")

The exported files of each hose biofilm were combined using excel and subsequent community analysis were conducted within the very same.

### NMDS - Community analysis based on Bray-Curtis Dissimilarity

```
ord_all <- ordinate(d.e.rare, method = "NMDS", distance = "bray")
#stress 0.03819
p_NMDS_all <- plot_ordination(physeq = d.e.rare, ordination = ord_all, shape = "Experiment", color="Orientation")

p_NMDS_all_number <- plot_ordination(physeq = d.e.rare, ordination = ord_all, shape = "Experiment", color="NumberCat")
```
OneTransp<-rgb(0.65098,0.38039,0.10196,max=1,alpha=0.5)
TwoTransp<-rgb(0.8745,0.76078,0.490196,alpha=0.25)
ThreeTransp<-rgb(0.50196,0.803921,0.44313725,alpha=0.25)
FourTransp<-rgb(0.00392,0.5215686,0.44313725,alpha=0.25)

One<-rgb(0.65098,0.38039,0.10196)
Two<-rgb(0.8745,0.76078,0.490196)
Three<-rgb(0.50196,0.803921,0.44313725)
Four<-rgb(0.00392,0.5215686,0.44313725)

plot(p_NMDS_all + geom_point(aes(color=(Orientation)),alpha=0.01,size=7) + theme_bw()+scale_shape_manual(values=c(15,16))+scale_color_manual(values=c(One,Two,Three,Four))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))
plot(p_NMDS_all_number  + geom_point(aes(color=(NumberCat)),alpha=0.01,size=7) + theme_bw() + scale_color_manual(values=c("gray5","gray10","gray15","gray20","gray25","gray30","gray35","gray40","gray45","gray50","gray55")+scale_shape_manual(values=c(15,16))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))
plot(p_NMDS_all_number  + geom_point(aes(color=(NumberCat)),alpha=0.01,size=7) + theme_bw()+scale_shape_manual(values=c(15,16))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))

BC_all<- phyloseq::distance(d.e.rare,method="bray")
SD_all<-data.frame(sample_data(d.e.rare))
adonis(BC_all~Experiment*Orientation,data=SD_all)

#Experiment explained 66% of all variation in communication, while orientation only explained 2%

#Permutation: free
#Number of permutations: 999

#Terms added sequentially (first to last)
#
#     Df   SumsOfSqs  MeanSqs   F.Model       R2  Pr(>F)  
#Experiment 1  52.111  52.111  793.730  0.65885 0.001 ***
#Orientation 2   1.706   0.853   12.991  0.02157 0.001 ***
#Residuals 385  25.277   0.066
#Total 388  79.094

#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Experiment explained 66% of all variation in communication, while orientation only explained 2%

BC_control<- phyloseq::distance(d.e.control,method="bray")
SD_control<-data.frame(sample_data(d.e.control))
adonis(BC_control~Orientation,data=SD_control)

#Permutation: free
#Number of permutations: 999

#Terms added sequentially (first to last)
#            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
#Orientation   1    1.3235  1.3235  54.928 0.21978  0.001 ***
#Residuals   195    4.6986  0.0241         0.78022
#Total       196    6.0221                 1.00000

#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

orientation could explain 22% of community variation.

adonis(BC_control~Orientation*NumberCat,data=SD_control)

#Permutation: free
#Number of permutations: 999

#Terms added sequentially (first to last)

#            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
#Orientation             1    1.3235 1.32353  84.328 0.21978
#NumberCat              10    1.6138 0.16138  10.282 0.26798
#Orientation:NumberCat  10    0.3381 0.03381   2.154 0.05615
#Residuals             175    2.7466 0.01570         0.45609
#Total                 196    6.0221                 1.00000

Pr(>F)
  #Orientation            0.001 ***
  #NumberCat              0.001 ***
  #Orientation:NumberCat  0.002 **
  #Residuals
  #Total

#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

____________

BC_real<- phyloseq::distance(d.e.real,method="bray")
SD_real<-data.frame(sample_data(d.e.real))
adonis(BC_real~Orientation,data=SD_real)

#Permutation: free
#Number of permutations: 999

#Terms added sequentially (first to last)

#            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
#Orientation   1    0.3827 0.38271  3.5336 0.01826  0.007 **
#Residuals   190   20.5779 0.10830         0.98174
#Total       191   20.9607                 1.00000

#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#orientation could only explain 1.8% of variation.

adonis(BC_real~Orientation*NumberCat,data=SD_real)
#Permutation: free
#Number of permutations: 999

#Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>SumsOfSqs</th>
<th>MeanSqs</th>
<th>F.Model</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation</td>
<td>1</td>
<td>0.3827</td>
<td>0.38271</td>
<td>8.9089</td>
<td>0.01826</td>
</tr>
<tr>
<td>NumberCat</td>
<td>10</td>
<td>11.8219</td>
<td>1.18219</td>
<td>27.5199</td>
<td>0.56400</td>
</tr>
<tr>
<td>Orientation:NumberCat</td>
<td>10</td>
<td>1.4533</td>
<td>0.14533</td>
<td>3.3831</td>
<td>0.06933</td>
</tr>
<tr>
<td>Residuals</td>
<td>170</td>
<td>7.3028</td>
<td>0.04296</td>
<td></td>
<td>0.34840</td>
</tr>
<tr>
<td>Total</td>
<td>191</td>
<td>20.9607</td>
<td>1.00000</td>
<td></td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Pr(>F)

#Orientation            0.001 ***
#NumberCat              0.001 ***
#Orientation:NumberCat  0.001 ***
#Residuals
#Total
#---
#Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

ord_real <- ordinate(d.e.real, method = "NMDS", distance = "bray")
# Run 20 stress 0.2258917
# ... New best solution
# ... Procrustes: rmse 0.00165251 max resid 0.02073589
# *** No convergence -- monoMDS stopping criteria:
#      20: stress ratio > sratmax

p3 <- plot_ordination(physeq = d.e.real, ordination = ord_real, shape = "Experiment", color = "Orientation")

p_NMDS_real <- plot_ordination(physeq = d.e.real, ordination = ord_real, color = "Orientation")

plot(p_NMDS_real + geom_point(aes(color=(Orientation)),alpha=0.01,size=7) + theme_bw()+scale_shape_manual(values=c(15))+scale_color_manual(values=c(One,Two,Three,Four))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))

p_NMDS_real_number <- plot_ordination(physeq = d.e.real, ordination = ord_real, shape = "Orientation", color="NumberCat")

plot(p_NMDS_real_number + geom_point(aes(color=(NumberCat)),alpha=0.01,size=7) + theme_bw() +scale_shape_manual(values=c(15,16))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))

ord_control <- ordinate(d.e.control, method = "NMDS", distance = "bray")
# Run 20 stress 0.2536307
# *** No convergence -- monoMDS stopping criteria:
#     1: no. of iterations >= maxit
#     19: stress ratio > sratmax
#p4 <- plot_ordination(physeq = d.e.control, ordination = ord_control, shape = "Experiment", color = "Orientation")
#p4

p_NMDS_control <- plot_ordination(physeq = d.e.control, ordination = ord_control, color = "Orientation", shape="Experiment")

plot(p_NMDS_control + geom_point(pch=15,aes(color=(Orientation)),alpha=0.01,size=7) + theme_bw()+scale_shape_manual(values=c(15))+scale_color_manual(values=c(Three,Four))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))

p_NMDS_control_number <- plot_ordination(physeq = d.e.control, ordination = ord_control, shape = "Orientation", color="NumberCat")

plot(p_NMDS_control_number + geom_point(aes(color=(NumberCat)),alpha=0.01,size=7) + theme_bw()+scale_shape_manual(values=c(15,16))+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+coord_fixed(ratio = 1))

d.n <- subset_samples(d, Character == "d")
d.n

# phyloseq-class experiment-level object
# otu_table() OTU Table: [ 545 taxa and 11 samples ]
# sample_data() Sample Data: [ 11 samples by 11 sample variables ]
# tax_table() Taxonomy Table: [ 545 taxa by 7 taxonomic ranks ]
# phy_tree() Phylogenetic Tree: [ 545 tips and 544 internal nodes ]
# refseq() DNAStringSet: [ 545 reference sequences ]

11 samples: 9 negative and 2 positive controls.

**Negativ controls**

There were two kinds of negative controls (NCs):

1. Controls taken to evaluate the potential of cross contamination during & between experimental processes, namely biofilm removal and needle sonication. These NCs are not necessarily expected to show zero or low counts.

2. Controls taken for potential contamination steps during the experiment, namely filtration for DNA analysis, process of DNA extraction, and process of PCR amplification. All of these samples are expected to show zero or low counts.
Negative sample description:
# NC-1-17    Control hose, Sample processing: Biofilm removal (brushing 10 mL of filtered water in petri dish), followed by needle sonication
  -> expectation = a decent number of counts, but acceptable for further analysis
# NC-1-18    Real hose, Sample processing: Biofilm removal, followed by needle sonication
  -> expectation = decent number of counts, but acceptable for further analysis
# NC-2       Real hose: Ultrapure water on sterile 0.2 µm filter
  -> expectation = zero (or low) counts
# NC-3       Control hose, Sample processing: Biofilm removal, followed by needle sonication
  -> expectation = decent number of counts, but acceptable for further analysis
# NC-4       Control hose: 0.2 µm filtered bottled Evian (used as matrix for biofilm suspension)
  -> expectation = zero (or low) counts
# NC-DNA     Real hose: sterile water extracted
  -> expectation = zero (or low) counts
# NC-Ext     Control hose: sterile water extracted
  -> expectation = zero (or low) counts (Real hose)
# NC-PCR-17  Control hose, PCR blanks -> expectation = zero (or low) counts
# NC-PCR-18  Real hose, PCR blanks -> expectation = zero (or low) counts

Negative control raw data was processed in the very same way as the real biofilm data. I.e., removing taxa with less than 2 reads and rarefaction to an even sampling depth of 25'000 reads/sample. Due to these restrictions, only 2 Negative controls showed positive reads for some taxa, the remaining controls were excluded during this process due to low total number of reads. By this, all Negative controls of the Real hose experiment, all DNA extraction controls, and PCR negative controls were excluded.

For the remaining Negative controls, the presence and relative abundance of the identified taxa (with relative abundance > 1% of the total number of reads) was compared to taxa identified in the corresponding data set (i.e, control hose biofilm data).
<table>
<thead>
<tr>
<th>ZOTU</th>
<th>Identity (Genus level)</th>
<th>Negative control &quot;NC-1-17&quot; (rel%)</th>
<th>Negative control &quot;NC-3&quot; (rel%)</th>
<th>Control hose biofilm (averaged rel%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>uncultured (Family: Cytophagaceae)</td>
<td>39.7</td>
<td>8.0</td>
<td>24.7</td>
</tr>
<tr>
<td>15</td>
<td>Bradyrhizobium</td>
<td>13.4</td>
<td>3.4</td>
<td>23.4</td>
</tr>
<tr>
<td>6</td>
<td>NA (Family: Comamonadaceae)</td>
<td>10.8</td>
<td>3.2</td>
<td>8.6</td>
</tr>
<tr>
<td>5</td>
<td>uncultured (Phylum: TM6)</td>
<td>10.4</td>
<td>5.9</td>
<td>9.6</td>
</tr>
<tr>
<td>7</td>
<td>NA (Family: Rhodobacteraceae)</td>
<td>3.3</td>
<td>2.2</td>
<td>6.3</td>
</tr>
<tr>
<td>814</td>
<td>NA (Family: P3OB-42)</td>
<td>2.6</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>63</td>
<td>Turneriella</td>
<td>2.2</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>187</td>
<td>Denitratisoma</td>
<td>1.8</td>
<td>1.5</td>
<td>5.7</td>
</tr>
<tr>
<td>466</td>
<td>Acinetobacter</td>
<td>1.5</td>
<td>-</td>
<td>0.0</td>
</tr>
<tr>
<td>14</td>
<td>Sediminibacterium</td>
<td>1.2</td>
<td>-</td>
<td>1.9</td>
</tr>
<tr>
<td>259</td>
<td>Acinetobacter</td>
<td>1.0</td>
<td>-</td>
<td>0.0</td>
</tr>
<tr>
<td>56</td>
<td>Pseudorhodoferax</td>
<td>-</td>
<td>67.1</td>
<td>0.0</td>
</tr>
<tr>
<td>714</td>
<td>Variovorax</td>
<td>-</td>
<td>1.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The majority of taxa identified in the Negative controls ("NC-1-17", "NC-3") were not dominant in the data of the control hose biofilm. However, for "NC-1-17", ZOTU3 and ZOTU15 showed a relative abundance of 39.7% and 13.4% respectively. Both these taxa were also amongst the most dominant ones in the Control hose biofilm community. In "NC-3", ZOTU3 showed a relative abundance of 7.7% and ZOTU5 one of 5.5%. As both these Negative controls were processing controls and the observed ZOTUs not present in the other tested negative controls (checked even though number of reads below 25'000 and thus excluded) it is supposedly a result of biofilm/DNA transfer between individual samples. It is therefore suggested to clean toothbrush heads in a way that also removes remaining DNA, e.g., with a chlorination step. For this study, transfer of biofilm material was minimized as good as possible (in the scope of the experiment). It is presumed, that the identified ZOTUs were dominant throughout the biofilm(s), as the contamination level would otherwise have been diluted out. Nevertheless, the high number of reads (i.e., relative abundance) for ZOTU3 in the Processing Negative controls ("NC-1-17" and "NC-3") cannot entirely be neglected and should therefore be taken into account when drawing conclusions on small-scale heterogeneity for this taxon.

**Positiv controls**

As a control for the PCR reaction and MiSeq sequencing, a positive control was amplified in addition to the actual samples (for both control and real hose separately). Here, a selfmade MOCK community (by the Genetic Diversity Center, ETH Zürich) was used, consisting of: *Burkholderia xenovorans*, *Bacillus subtilis*, *Escherichia coli*, *Micrococcus luteus*, *Pseudomonas protegens*, *Paenibacillus sabinae*, and *Streptomyces violaceoruber*. The success of their amplification was verified, following the same data processing steps as for the Negative controls (see above).
The presence and abundance of each Positive control taxon was checked and compared between the sequencing run of both Control and Real hose.

<table>
<thead>
<tr>
<th>ZOTU</th>
<th>Identity (Genus level)</th>
<th>Control hose biofilm (rel%)</th>
<th>Real hose biofilm (rel%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>Burholderia-Paraburkholderia</td>
<td>41.62</td>
<td>34.19</td>
</tr>
<tr>
<td>68</td>
<td>Bacillus</td>
<td>49.83</td>
<td>55.24</td>
</tr>
<tr>
<td>122</td>
<td>Escherichia-Shigella</td>
<td>4.04</td>
<td>3.72</td>
</tr>
<tr>
<td>121</td>
<td>Micrococcus</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>347</td>
<td>Pseudomonas</td>
<td>0.36</td>
<td>0.33</td>
</tr>
<tr>
<td>808</td>
<td>Paenibacillus</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>499</td>
<td>Streptomyces</td>
<td>0.06</td>
<td>0.03</td>
</tr>
</tbody>
</table>

All taxa that were supposed to be identified within the amplified and sequenced MOCK community were detectable (not on species but genus level). As the positive control was run for both control and real hose biofilms, the amplification success could be compared between the two sequencing runs.

In the table above, the relative percentages of the members of the MOCK community are displayed (rarefied to an even depth of 25'000 reads/sample to allow for comparability to the main experimental samples). As the relative % of the genera is comparable between the sequencing of control and real hose biofilms, the amplification of the Positive control can be evaluated as successful.

**Replicate sequencing of specific samples**

For the evaluation of the sequencing data, some samples have been amplified and sequenced in replicates. Here, 1 sample of the control hose was amplified in triplicate ("U093") as well as 6 samples of the real hose ("B005", "B013", "B027", "B042", "B050", "B082").

**Subset data set**

```r
# REPLICATE SAMPLES FOR AMPLIFICATION CONTROL
d.t <- subset_samples(d, Kind == "t")
d.t
```

---

**Remove ZOTUs with zero counts and singletons**

Samples with low counts might not represent diversity very well and should therefore be removed.
# DETERMINE NUMBER OF ZOTUs WITH <1 AND/OR =1 READS

```r
sum(taxa_sums(d.t) < 1)
[1] 252
```

```r
sum(taxa_sums(d.t) == 1)
[1] 25
```

d.t.noZ <- prune_taxa(taxa_sums(d.t) > 1, d.t)
d.t.noZ
```

# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 268 taxa and 20 samples ]
# sample_data() Sample Data:       [ 20 samples by 11 sample variables ]
# tax_table()   Taxonomy Table:    [ 268 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 268 tips and 267 internal nodes ]
# refseq()      DNAStringSet:      [ 268 reference sequences ]
```

Set minimum depth and seed
# CHECK MINIMAL DEPTH
min_depth <- min(sample_sums(d.e))
min_depth

# [1] 5648

# CHANGE MINIMAL DEPTH (based on read-count distribution)
min_depth <- c(25000)
min_depth

# [1] 25000

# SET SEED AND RAREFY EVEN DEPTH
set.seed(1)
d.t.rare <- rarefy_even_depth(d.t.noZ, sample.size = min_depth, replace = TRUE, rngseed = 1)

# `set.seed(1)` was used to initialize repeatable random subsampling.
# Please record this for your records so others can reproduce.
# Try `set.seed(1); .Random.seed` for the full vector
# ...
# 2 samples removed because they contained fewer reads than `sample.size`.
# Up to first five removed samples are:
# B005-2B005-3
# ...
# 240OTUs were removed because they are no longer
# present in any sample after random subsampling

d.t.rare

# phyloseq-class experiment-level object
# `otu_table()` OTU Table: [ 244 taxa and 18 samples ]
# `sample_data()` Sample Data: [ 18 samples by 11 sample variables ]
# `tax_table()` Taxonomy Table: [ 244 taxa by 7 taxonomic ranks ]
# `phy_tree()` Phylogenetic Tree: [ 244 tips and 243 internal nodes ]
# `refseq()` DNAStringSet: [ 244 reference sequences ]

Export for final analysis of sequencing data outside the R environment

# EXPORT FOR ALL REPLICATE SAMPLES
ot3 <- as(otu_table(d.t.rare), 'matrix')
tax3 <- as(tax_table(d.t.rare), 'matrix')
write.csv(ot3, file = "Lineage_Replicates_1.csv")
write.csv(tax3, file = "Lineage_Replicates_2.csv")

The exported files of each hose biofilm were combined using excel and subsequent community analysis were conducted with the very same.

Sample "B005" could not be used for replicate analysis as 2 out of 3 samples were removed during rarefaction. Also, "B-027" was removed from further analysis as only 2 replicates remained.
<table>
<thead>
<tr>
<th>10 most dominant ZOTUs</th>
<th>average (B013)</th>
<th>stdev (+/-)</th>
<th>average (B042)</th>
<th>stdev (+/-)</th>
<th>average (B050)</th>
<th>stdev (+/-)</th>
<th>average (U082)</th>
<th>stdev (+/-)</th>
<th>average (U093)</th>
<th>stdev (+/-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>15'338</td>
<td>200</td>
<td>10'005</td>
<td>253</td>
<td>14'288</td>
<td>344</td>
<td>8'254</td>
<td>22</td>
<td>8'946</td>
<td>250</td>
</tr>
<tr>
<td>#2</td>
<td>4'436</td>
<td>133</td>
<td>6'821</td>
<td>59</td>
<td>3'565</td>
<td>64</td>
<td>6'338</td>
<td>12</td>
<td>3'883</td>
<td>116</td>
</tr>
<tr>
<td>#3</td>
<td>2'051</td>
<td>63</td>
<td>1'594</td>
<td>99</td>
<td>2'124</td>
<td>117</td>
<td>6'234</td>
<td>91</td>
<td>3'001</td>
<td>94</td>
</tr>
<tr>
<td>#4</td>
<td>1'394</td>
<td>25</td>
<td>1'354</td>
<td>50</td>
<td>1'413</td>
<td>73</td>
<td>1'172</td>
<td>17</td>
<td>2'178</td>
<td>68</td>
</tr>
<tr>
<td>#5</td>
<td>309</td>
<td>37</td>
<td>1'033</td>
<td>29</td>
<td>579</td>
<td>19</td>
<td>948</td>
<td>80</td>
<td>1'406</td>
<td>39</td>
</tr>
<tr>
<td>#6</td>
<td>288</td>
<td>23</td>
<td>613</td>
<td>22</td>
<td>526</td>
<td>21</td>
<td>260</td>
<td>22</td>
<td>1'404</td>
<td>76</td>
</tr>
<tr>
<td>#7</td>
<td>229</td>
<td>25</td>
<td>547</td>
<td>27</td>
<td>338</td>
<td>38</td>
<td>245</td>
<td>8</td>
<td>1'214</td>
<td>27</td>
</tr>
<tr>
<td>#8</td>
<td>130</td>
<td>5</td>
<td>422</td>
<td>18</td>
<td>290</td>
<td>18</td>
<td>228</td>
<td>20</td>
<td>427</td>
<td>16</td>
</tr>
<tr>
<td>#9</td>
<td>119</td>
<td>7</td>
<td>413</td>
<td>41</td>
<td>258</td>
<td>29</td>
<td>103</td>
<td>10</td>
<td>401</td>
<td>43</td>
</tr>
<tr>
<td>#10</td>
<td>81</td>
<td>12</td>
<td>245</td>
<td>22</td>
<td>233</td>
<td>14</td>
<td>101</td>
<td>20</td>
<td>204</td>
<td>6</td>
</tr>
</tbody>
</table>

Percentages of standard deviations in the total number of reads varied between 0.2 - 19.7 % (average: 5.4 ± 4.0%, n = 50), considering the ten most dominant ZOTUs of individual (triplicate) samples.