RUPRECHT-KARLS-UNIVERSITY HEIDELBERG
FACULTY OF BIOSCIENCES
MASTER OF SCIENCE IN MOLECULAR BIOTECHNOLOGY

Comparing Metabarcoding Analysis Methodologies for Nanopore Sequencing in Clinical Application

---

A MASTER THESIS *describing the work to be performed in the*

Bioinformatics and Omics Data Analytics Group
Functional and Structural Genomics
German Cancer Research Center

in the time from September 2019 *to* April 2020

Author:
Marc RÜBSAM

Born:
16th of June 1992
Erlabrunn, Germany

First Referee:
Dr. Matthias SCHLESNER

Second Referee:
PD Dr. med. Niels HALAMA

Declaration of Authorship

The work presented in this Master Thesis was performed in the Bioinformatics and Omics Data Analytics Group at the German Cancer Research Center in the time from September 2019 to April 2020 to complete the Master's program Molecular Biotechnology at the Faculty of Biosciences at the University of Heidelberg.

First Referee:	Dr. Matthias SCHLESNER Functional and Structural Genomics German Cancer Research Center
Second Referee:	PD Dr. med. Niels HALAMA Tumor Immunology Nationales Centrum für Tumorerkrankungen Heidelberg

---

I Marc RÜBSAM herewith declare that:

- I wrote this Master Thesis independently under supervision and that I used no other sources and supporting materials than those indicated;
- the adoption of quotation from the literature/internet as well as thoughts from other authors were indicated in the thesis;
- my Master Thesis was not submitted to any other examination.

I am aware of the fact that a false declaration will have legal consequences.

Heidelberg, 04.04.2020 _____

Place and Date Signature Marc RÜBSAM

Abstract

Comparing Metabarcoding Analysis Methodologies for Nanopore Sequencing in Clinical Application

The analysis of microbial marker genes like the 16S rRNA gene is an efficient way to study the composition of microbial communities. However, the limited read length of second-generation sequencing prevents the analysis of the full set of variable regions in the marker gene. Nanopore sequencing, a third-generation technique, can be used to sequence the near-full-length 16S rRNA gene, but the lack of automated analysis pipelines limits its utilization in a clinical study such as the PROMISE trial. We have therefore developed a novel pipeline, called MeBaPiNa. The implemention and validation of commonly use techniques revealed an clear superiority of the utilization of individual reads over feature extraction methods. In a showcase of samples of different sources, we were able to analyze the taxonomic composition of the samples and detect patient and time point specific variability.

Zusammenfassung

Vergleich von Metabarcoding-Analysemthoden zur Auswertung von Nanopore-Sequenzierungsdaten in der klinischen Anwendung

Die Analyse von mikrobiellen Markergenen wie dem 16S rRNA-Gen ist eine effiziente Methode, um die Zusammensetzung von Mikrobengemeinschaften zu untersuchen. Die begrenzte Read-Länge der Sequenzierung der zweiten Generation verhindert jedoch, die Analyse des vollständigen Satzes aller variabler Regionen im Markergen. Die Nanoporen-Sequenzierung, eine Technik der dritten Generation, kann zur Sequenzierung des nahezu vollständigen 16S rRNA-Gens verwirklichenn, aber das Fehlen automatisierter Analysepipelines schränkt seine Verwendung im Rahmen einer klinischen Studie, wie der PROMISE-Studie, ein. Wir haben daher eine neue Pipeline entwickelt, die MeBaPiNa genannt wird. Die Implementierung und Validierung allgemein gebräuchlicher Techniken zeigte eine klare Überlegenheit der Verwendung von individuellen Reads gegenüber den Methoden zur Feature-Extraktion. In einem exemplarischem Beispielset von Proben unterschiedlicher Herkunft analysierten wir die taxonomische Zusammensetzung der und stellten patienten- und zeitpunktspezifische Variabilität fest.

Acknowledgemnent

I would like to thank Dr. Matthias SCHLESNER and PD Dr. med. Niels HALAMA for offering the possibility to take part in the work of their research groups and giving advice and suggestions promoting the project and supervising the thesis work. The same gratitude goes to Dr. med. Silke GRAULING-HALAMA for providing the input data required for this work, giving valuable advise for the analysis of the taxonomic classification results and involving me in the design process of the experiments. Further, I would like to thank Pornpimol CHAROENTONG, PhD for the personal supervision sharing experience and giving feedback and advise. I would also like to thank Daniel BROWN who has provided the computational environment needed for the analysis. Finally, I would also like to thank my brother Jan RÜBSAM for the effort he put into the proofreading of this thesis.

Contents

List of Figures

List of Tables

Abbreviations

16S rRNA	16 Svedberg small-subunit ribosomal RNA
23S rRNA	23 Svedberg small-subunit ribosomal RNA
ASV	Amplicon sequence variants
EC	Extraction control
FTP	File transfere protocol
gDNA	Genomic DNA
Gb	Giga bases or giga base pairs
kb	Kilo bases or kilo base pairs
Mb	Mega bases or mega base pairs
MeBaPiNa	Metabarcoding analysis pipeline for Nanopore datasets
NSCLC	Non-small-cell lung carcinoma
NTC	No-template control
ONT	Oxford Nanopore Technology
OTU	Operational taxonomic units
rrn operon	Ribosomal RNA operon
rRNA	Ribosomal RNA
ZyDNA	ZymoBIOMICS™ Microbial Community DNA Standard
ZyCell	ZymoBIOMICS™ Microbial Community Standard

1 Introduction

1.1 Nanopore Sequencing

Long-read sequencing is considered the third generation of sequencing techniques as it stands in contrast to the short reads produced by second-generation sequencing and some of the limitations associated with them (Jain et al. 2016; van Dijk et al. 2018). It is capable of producing reads with a length of 2 Mb (Payne et al. 2018) from a single template molecule in real time (Jain et al. 2018; Rang et al. 2018).

The long-read sequencing technique developed by ONT (Oxford Nanopore Technologies) is referred to as Nanopore sequencing. In contrast to other sequencing platforms, Nanopore sequencing does not utilize DNA synthesis to infer the composition of the template strand. Instead, the strand is fed through a name-giving pore inside an unpolar membrane across which a voltage potential is applied (de Lannoy et al. 2017). The molecule, traversing the pore, interferes with the ion flow through the pore, which results in detectable disturbance of the otherwise constant current flow. The strength of the induced fluctuation depends on the properties of the molecule inside the pore and by tracking the current changes over time, the composition of the molecule, i.e. its nucleotide sequence, can be inferred (Rang et al. 2018; Wick et al. 2019). A standard flow cell of ONTs MinION sequencer consists of 512 current channels, each with four wells containing one pore per well for a total of 2048 pores per flow cell (de Lannoy et al. 2017). Hence, many template molecules can be sequenced in parallel facilitating a throughput of 15-30 Gb in a 48 h run (Oxford Nanopore Technologies, Limited. 2019b).

The independence from fluorescence measurements and large reagent flow controls also results in an drastic reduction in spatial dimension as well as initial capital investment of the Nanopore sequencing platforms. The MinION sequencer weights only about 90 g and is comparable in size to an office stapler (Rang et al. 2018). This portability facilitates sequencing in a wide range of use cases and locations unfeasible for other platforms (Castro-Wallace et al. 2017; Edwards et al. 2016;

Hoenen et al. 2016). Even so, its affordability, scalability and fast turn-around time make it an appealing alternative to second-generation platforms in clinical and laboratory applications as well.

1.2 Analysis of the Human Microbiome

The microorganisms populating the human body are essential for a normal body function. However, alterations in the microbiota are associated with multiple disease conditions, sometimes liked to changes of an individual species (reviewed in Falony et al. 2019; Pflughoeft and Versalovic 2012; Wang et al. 2017, among others). As such, tools providing a high resolution of the microbial composition are needed to detect changes between different patient cohorts or time points and extract clinical markers or identify pathogens.

Cultivation remains the method of choice in many clinical applications, as it easily provides a large number of clonal cells for classification and testing. However, a cultivation is difficult or impossible for many organisms, leading to biased observations (Hugerth and Andersson 2017).

1.2.1 16S ribosomal RNA Metabarcoding

The limitations of cultivation can be overcome by analyzing the composition of the microbial communities at the genome level. This microbiome can be extracted directly from the samples and, in an ideal scenario, represents all species independent of their growth conditions.

Instead of analyzing the full complexity of the microbiome, it is sufficient to investigate a genomic region or gene with enough genomic distance between the species to be able to discriminate them. The prokaryotic 16S and 23S small-subunit rRNA (ribosomal RNA) genes are good candidates for such a marker. They are orthologous genes universally expressed in all known prokaryotes and show negelectable horizontal gene transfer (Daubin et al. 2003; Hugerth and Andersson 2017). At the same time, they contain a mixture of highly conserved regions and less-conserved variable regions usable for taxonomic classification and inference of evolutionary relationships (Woese and Fox 1977; Yang et al. 2016), as

illustrated for the 16S rRNA gene in Figure 1. The conserved regions facilitate the analysis of microbial communities in a highly parallel manner as primers aligning to these regions can be used to extract parts of the marker gene and analyze the intermediary variable regions by sequencing (Klindworth et al. 2013).

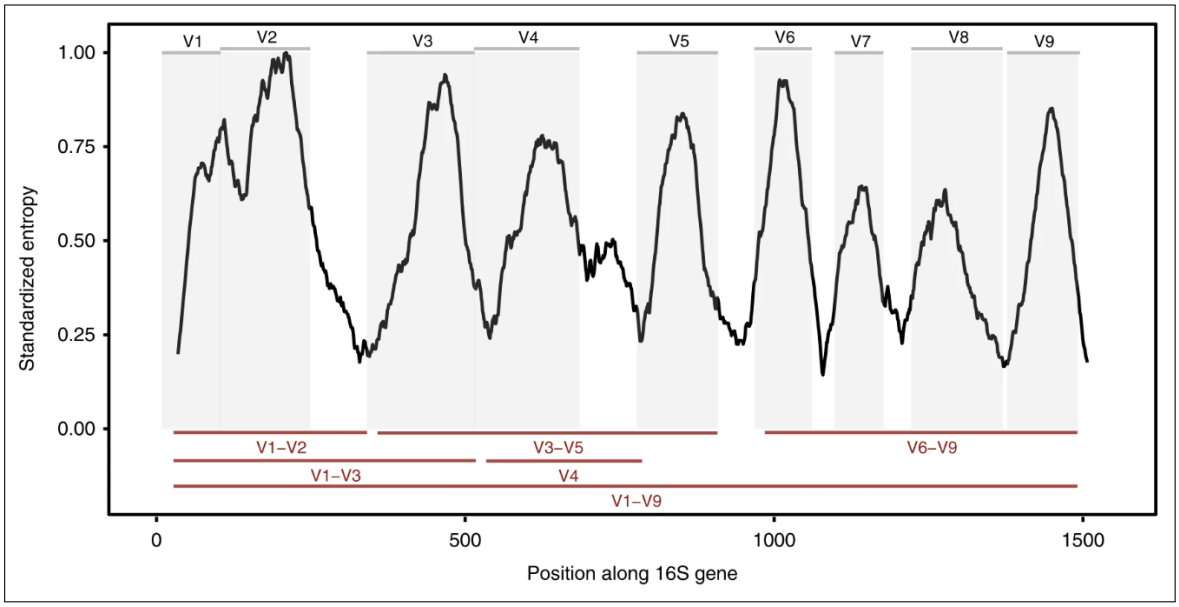

FIGURE 1. Constant and variable 16S-rRNA-gene regions: Shown is a graphical illustration of the variability in the 16S rRNA gene from Johnson et al. (2019). The black line shows the amount of variability between species along the 16S rRNA gene. grey bars highlight the variable regions V1-V9. Red lines below show gene segments used for metabarcoding analysis. V1-V9 can not be fully sequenced using second-generation sequencing.

1.2.2 Metabarcoding using Nanopore Sequencing

The full rrn (ribosomal RNA) operon sequence covers over 5 kb, and contains the 16S rRNA gene (1'522 bp), an internal transcribed spacer (820-1'100 bp) and the 23S rRNA gene (2'971 bp) (Bouchet et al. 2008; Feibelman et al. 1994).

The maximum combined read length of common second-generation platforms reaches up to 600 bp (Derakhshani et al. 2015), which is to short to cover a full rrn operon or even a full rRNA gene (Klindworth et al. 2013; Yang et al. 2016). Hence, the taxonomic classification with second-generation platforms are limited to sequencing a subset of variable regions of the marker genes. This, however,

reduces the accuracy and resolution of abundance prediction and underestimates the microbial diversity (Myer et al. 2016; Shin et al. 2016).

As described in Section 1.1, Nanopore sequencing achieves much greater read length and can be used to sequence amplicons of the near full-length 16S rRNA gene (see V1-V9 segment in Figure 1). This way Nanopore sequencing was successfully applied to study mock communities (Benítez-Páez et al. 2016; Cuscó et al. 2019; Cuscó et al. 2017; Kai et al. 2019; Mitsuhashi et al. 2017) and single strain isolates (Kai et al. 2019; Ma et al. 2017), cryonite samples (Edwards et al. 2016), seawater (Curren et al. 2019) and algae-associated communities (Shin et al. 2018), a wastewater microbiome (Ma et al. 2017), dog skin microbial communities (Cuscó et al. 2017) and the mouse gut microbiota (Shin et al. 2016) as well as in clinical pathogen determination (Mitsuhashi et al. 2017; Moon et al. 2017).

1.3 Taxonomic Classification Approaches

Despite the longstanding background of 16S sequencing (starting as early as 1977 with Woese and Fox 1977), no consent has been found in the way 16S sequencing data is evaluated as new methods are constantly developed and benchmark results disagree (Bolyen et al. 2019; Callahan et al. 2016; Edgar 2017, 2018; Hugerth and Andersson 2017) and reviews focus on methodological comparison (Hugerth and Andersson 2017; Malla et al. 2018; Pollock et al. 2018). For long-read sequencing, this situation is even more complicated as most methods have been developed for short-read sequencing and not all assumptions are transferable to long-read sequencing datasets due to its increased error rate of about 7.5% (Jain et al. 2017). It is therefore not surprising that the publications listed in Section 1.2.2 use a variety of different analytical approaches, which makes comparison difficult. The investigation of differences, based on the processing of the same data set, is necessary for a direct and meaningful comparison of the methodologies.

1.3.1 Feature Extraction

The analysis methods developed for second-generation sequencing focus on the high redundancy of the reads generated by the sequencers. Some of these reads contain sequencing errors or represent artifacts created during library preparation.

The algorithms try to cluster similar reads into features and extract their centroid sequences, to reduce the noise and with it unnecessary complexity in the dataset (Bolyen et al. 2019; Hugerth and Andersson 2017).

In the OTU (operational taxonomic unit) picking method, the reads are clustered by an identity threshold in comparison to a set of reference sequences or to each other (Edgar 2013; Rognes et al. 2016). This approach has been criticized for the arbitrary selection of the identity threshold (most often 97%), which can be above the actual genomic distance between the organisms in a community leading to an artificial decrease in richness and diversity (Callahan et al. 2016; Edgar 2017). Because of the increased error rate in long-read sequencing, this approach has been shown to be unsuccessful (Ma et al. 2017) or required a even stronger reduction of the identity threshold (a successful clustering at 85% was shown by Curren et al. 2019).

Because of its disadvantages, OTU-picking has been superseded by techniques trying to recover true amplicon sequence variants (ASV) by analyzing the introduced errors and clustering reads considered identical under the inferred error model (Callahan et al. 2016; Edgar 2016a). However, despite the advantage this shows for second-generation sequencing, there are no publications about its application on long-read sequencing.

1.3.2 Classification of Features or Reads

Independent of the procedure, the extracted features have to be assigned to taxa in order to analyze the composition of the microbial community and apply further analyses. For second-generation sequencing data, this is most commonly achieved using classifiers based on a naïve Bayes approach (Pedregosa et al. 2011; Wang et al. 2007) or k-mer mapping (Edgar 2016b; Wood et al. 2019).

Since the applicability of feature extraction for long-read sequencing is limited, the additional information content from its extended read length has been successfully utilized to adapt the classifiers to use individual input reads as features (Benítez-Páez et al. 2016; Edwards et al. 2016; Kerkhof et al. 2017; Ma et al. 2017; Mitsuhashi et al. 2017). Further, the increased read-length also facilitates approaches considered inferior for short reads. The full-length alignment of long reads to a reference sequence database, deemed to create high numbers of false positives in second-generation sequencing, has been used to accurately identify

species in a variety of samples (Cuscó et al. 2019; Cuscó et al. 2017; Kai et al. 2019; Li et al. 2016; Shin et al. 2018, 2016), but not compared to the results of established approaches.

1.4 Analysis Pipeline

Pipeline implementations of the tools for feature extraction and taxonomic classification of second-generation sequencing have already been developed (Bolyen et al. 2019; Minot et al. 2015; Schloss et al. 2009). However, with the exception of the cloud-based EPI2ME platform from ONT (Oxford Nanopore Technologies, Limited. 2019c), non offers full support for long-read sequencing and full-length alignment approaches are not integrated. Unfortunately EPI2ME is not configurable and the exact underlying methodology is not disclosed by ONT, affecting the reproducibility of its results.

Hence, a pipeline for the metabarcoding analysis for Nanopore sequencing datasets has to be developed, which:

- includes the four conceptual methodologies of taxonomic classification
  - naïve Bayes or k-mer mapping classification of OTUs
  - naïve Bayes or k-mer mapping classification of ASVs
  - naïve Bayes or k-mer mapping classification of individual reads
  - full-length alignment of individual reads
- is configurable in terms of methodology and its parameters
- is reproducible

In the work presented here, we aim to design a pipeline that meets the requirements listed above. Further, in clinical applications a minimal user input during sample analysis is desired. We therefore aim to fulfill additional requirements to:

- include basecalling and quality filtering of the raw sequencing reads
- include output visualization and statistics
- automatically perform the steps while considering dependencies
- have the ability to rerun certain steps
- process multiple files in parallel
- use free and maintained software

1.5 Samples

The 16S rRNA gene has been widely used to analyze the human microbiome with second-generation sequencing (Huttenhower et al. 2012), but similar results for Nanopore sequencing are sparse, despite its successful application in many use cases (see Section 1.2.2).

Here we utilize a set of sequencing datasets from mock community samples compare the methodologies used for taxonomic classification of metabarcoding datasets. The method with the highest resolution is then used characterize a sequencing datasets from clinical samples in the context of the PROMISE trial.

1.5.1 Mock Community

Mock communities of known composition are a valuable tool to analyze the performance of novel tools or methodologies for analyses of microbial communities. The known ratios of community members in the input can be used to calculate the deviation of the determined composition to this reference (Pollock et al. 2018). Further, the mixture of gram-positive and -negative species in a cell based community helps to detect extraction biases and variable genomic GC-contents can be used to analyze PCR biases (Browne et al. 2020; Gorzelak et al. 2015; McOrist et al. 2002; Walker et al. 2015).

1.5.2 PROMISE Trial

As described in Section 1.2, the microbiome is associated with many disease conditions. One of these conditions is cancer, where microorganism have been identified as driving factors influencing the disease progression (recently reviewed in Helmink et al. 2019).

The performed work is part of the PROMISE trial (PROMISE trial consortium 2018). This clinical trial will recruit 150 newly diagnosed stage IV NSCLC (Non-small-cell lung carcinoma) patients and aims to assess the predictability of outcome based on immunological signatures in lung cancer. The analyzed biomarker signatures include tissue biopsies and stool samples from several time points during the treatment, including initial diagnosis (T1/T2, stool and tissue samples), 6-10 weeks

after treatment initiation (T2ER, stool samples) and after first (T3, stool and tissue samples) and second disease progression (T4, tissue samples).

Data from enrolled trial patients as well as other clinical sequencing datasets is used to showcase the ability of the pipeline to classify the microbial composition of multiple sample sources, reflecting the trial material, and detect changes over time.

1.5.3 Control Samples and Kit Contaminations

The amount of microbial materials varies between sample sources and tissue samples, especially from lung tissue, tend to have low amounts of microbial content (O'Dwyer et al. 2016; Savage 1977; Sender et al. 2016) and lower amount of available material, compared to stool samples. Hence, the number of cycles of the 16S rRNA gene amplification have to be increased to achieve high enough concentrations for sequencing.

All reagents used during the preparation of samples and library are know to contain small amounts of contaminations in form of bacterial DNA. Whilst being mostly unproblematic for stool samples, the increased cycle count for low biomass samples strongly amplifies this contamination to the extend where it comprises a large part of the input library (Salter et al. 2014; Weiss et al. 2014). Extraction and no-template controls have to be used to detect such contaminations and allow for an selective filtering.

2 Results

2.1 Sequence Generation

2.1.1 Sample Set

The pipeline was build to aid the analysis of microbial communities as part of the PROMISE trial (see Section 1.5.2). We included samples of different sources to represent these microbiomes during the development process. Specifically, we were able to include a mock community as well as a set of clinical samples of stool and tissue. The set of samples is specified in Table 1.

Stool samples of six individual patients were used to investigate the gut microbiome. Five of these samples originated from the PROMISE trial itself and timepoints before and after treatment initiation were compared for two of the patients to investigate the ability to track changes in the gut microbiome during treatment. Another four individual patient samples from two different tissues of origin, three lung and one ovary, were processed to analyze the tissue microbiome. A comparison of extracted and enriched DNA was included for the lung samples. The stool and tissue samples were sequenced in separate runs. The mock community in form of purified DNA or whole cells was sequenced in three instances across both runs and allows us to determine reproducibility over the sequencing runs as well as sensitivity and specificity of the different analysis methodologies.

2.1.2 Library Preparation and Nanopore Sequencing

The set of samples listed in Table 1, was used to produce the sequencing reads on which basis the performance of different analysis methodologies was investigated to develop an automated pipeline. DNA extraction and enrichment as well as library preparation and sequencing was performed in parallel to my work by Dr. med. Silke GRAULING-HALAMA. Two individual runs were performed. The first run (I) was focused on the gut microbiome including several stool samples and both sources

TABLE 1. Sample cohort: An overview of the samples sequenced during two individual runs. Each sample ID is defined by the material source and a patient identifier. Samples from the PROMISE trial (marked by '*') are further differentiated by the time point they have been collected. DNA extraction, enriched for microbial DNA and library preparation together with sequencing was performed in batches as indicated. Barcodes were added during library preparation and samples were multiplexed. Samples of the ZymoBIOMICS microbial community DNA standard (ZyDNA) and the ZymoBIOMICS whole cell mock community standard (ZyCell) were included. Each run includes a no-template control (NTC) and an extraction control (EC). Steps described here were performed by Dr. med. Silke GRAULING-HALAMA.

ID	source	patient	timep.	extrac.	enrich.	run	barc.
NTC	-	-	-	-	-	I	01
EC	-	-	-	$III/1$	-	I	02
ZyDNA	DNA	-	-	-	-	I	03
ZyCell	cells	-	-	$IV/2$	-	I	04
St61	stool	OP61	-	$IV/3$	-	I	05
St01-1*	stool	1-001	T1T2	$III/2$	-	I	06
St02-1*	stool	1-002	T1T2	$IV/4$	-	I	07
St01-2*	stool	1-001	T2ER	$III/4$	-	I	08
St03-1*	stool	1-003	T1T2	$III/5$	-	I	09
St04-1*	stool	1-004	T1T2	$III/6$	-	I	10
St05-1*	stool	1-005	T1T2	$IV/5$	-	I	11
St02-2*	stool	1-002	T2ER	$III/8$	-	I	12
NTC	-	-	-	-	-	II	01
EC	-	-	-	$II/1$	-	II	02
ZyDNA	DNA	-	-	-	-	II	03
Lu05	lung	Lu5	-	$II/2$	-	II	04
Lu13	lung	Lu13	-	$II/3$	-	II	05
Lu18	lung	Lu18	-	$II/4$	-	II	06
EC-en	-	-	-	$II/1$	II	II	07
Lu05-en	lung	Lu5	-	$II/2$	II	II	08
Lu13-en	lung	Lu13	-	$II/3$	II	II	09
Lu18-en	lung	Lu18	-	$II/4$	II	II	10
Lu05-en2	lung	Lu5	-	I	I	II	11
Ov85-en	ovary	Ov85	-	I	I	II	12

of the mock community. The second run (II) was focused on tissue microbiomes and included primary and enriched DNA from different tissue samples as well as an additional mock community sample.

The duration of run I was 48 h, which resulted in a throughput of 11.9*10^6 raw reads

(see Table 2). Based on the high throughput of the first run, run II was stopped after 19 h yielding 8.5*10^6 raw reads. The resulting ratio of the run durations between run I and II was approximately 1:2.5, but the ratio of produced reads was only about 1:1.4.

TABLE 2. Sequencing throughput: Raw read count of the sequencing runs and determining experimental factors. Namely, the number of available pores reported during the flow cell check, the total mass of DNA pooled during library preparation (and averaged over the number of non-control samples, see Table 15) and run duration.

run	reads [10^6]	pores	mass (av.) [ng]	duration [h]
I	11.9	1370	83.0 (8.3)	48
II	8.5	1464	96.8 (8.9)	19

It has to be noted that the cumulative input DNA mass for run II was slightly higher than in run I as the extraction and no-template control of run II showed an amplification caused by contaminations in the kit (data not shown). However, the input mass per non-control sample was comparable between both runs (see Table 2). Both runs were performed on fresh flow cells with comparable numbers of pores. No abnormalities in the distribution of active pores or reads per pore were detected during the runs (see Figure S1), excluding the flow cells as source of the differences in throughput. Similarly, a decrease of active pores and throughput over time was observed for both flow cells (Figure S2), as the running buffer is depleted and well membranes get distorted. In the respective time frame, the number of active pores and throughput was lower in the first run, starting with about 30% less channels and 50% less reads. However, the decrease over time was lower for run I as well, reducing the difference to about 20% less channels and 45% less reads after 19 h run time.

Despite the differences, both sequencing runs were successfully completed and resulted in a high number of reads in the raw signal format, suitable for downstream processing.

2.2 Metabarcoding Analysis Pipeline

The central aim was the creation of a configurable pipeline based on available and maintained software packages for automated analysis of the raw sequencing reads.

The full workflow is illustrated in Figure 2.

The MeBaPiNa (metabarcoding analysis pipeline for Nanopore datasets) called pipeline was implemented in snakemake. This workflow management system facilitates the connection of the individual steps illustrated in Figure 2 into a fully automated pipeline. An Excel sheet, containing the information listed in Table 1, was provided to the pipeline as input. The list of samples to be analyzed, the desired analysis methodology and additional parameter were provided in a separate configuration file. The pipeline was designed to automatically create result files, figures and statistics at multiple points of the analysis and thus offers an automated reads-to-report pipeline.

Upon execution, the pipeline automatically selects the raw read files of the run containing the specified samples and determined the required set of rules and their dependency and executed them in the correct order. It first initiated the basecalling and quality-control of the raw reads to obtain their nucleotide sequences and remove artifacts and low quality reads. This is followed by one of the parallel analysis branches for feature extraction and taxonomic classification, depending on the methodology specified in the configuration file. The exact procedure of the analysis steps will be discussed in the context of the results in the corresponding sections of this manuscript.

2.3 Quality Control

2.3.1 Basecalling

The raw read files had to be converted into nucleotide sequences required for quality control and downstream analysis. MeBaPiNa automatically selected the raw read files of the run containing the desired sample set and executed the basecalling rule build around the guppy basecaller developed by ONT. The guppy basecaller is based on a recurrent neural network capable of converting the raw current track into a nucleotide sequence.

The basecalling results for both runs are listed in Table 3. The total number of reads was not influenced by the basecalling process (compare Table 2) which translated into a total of 18.2 Gb for run I and 8.3 Gb for run II.

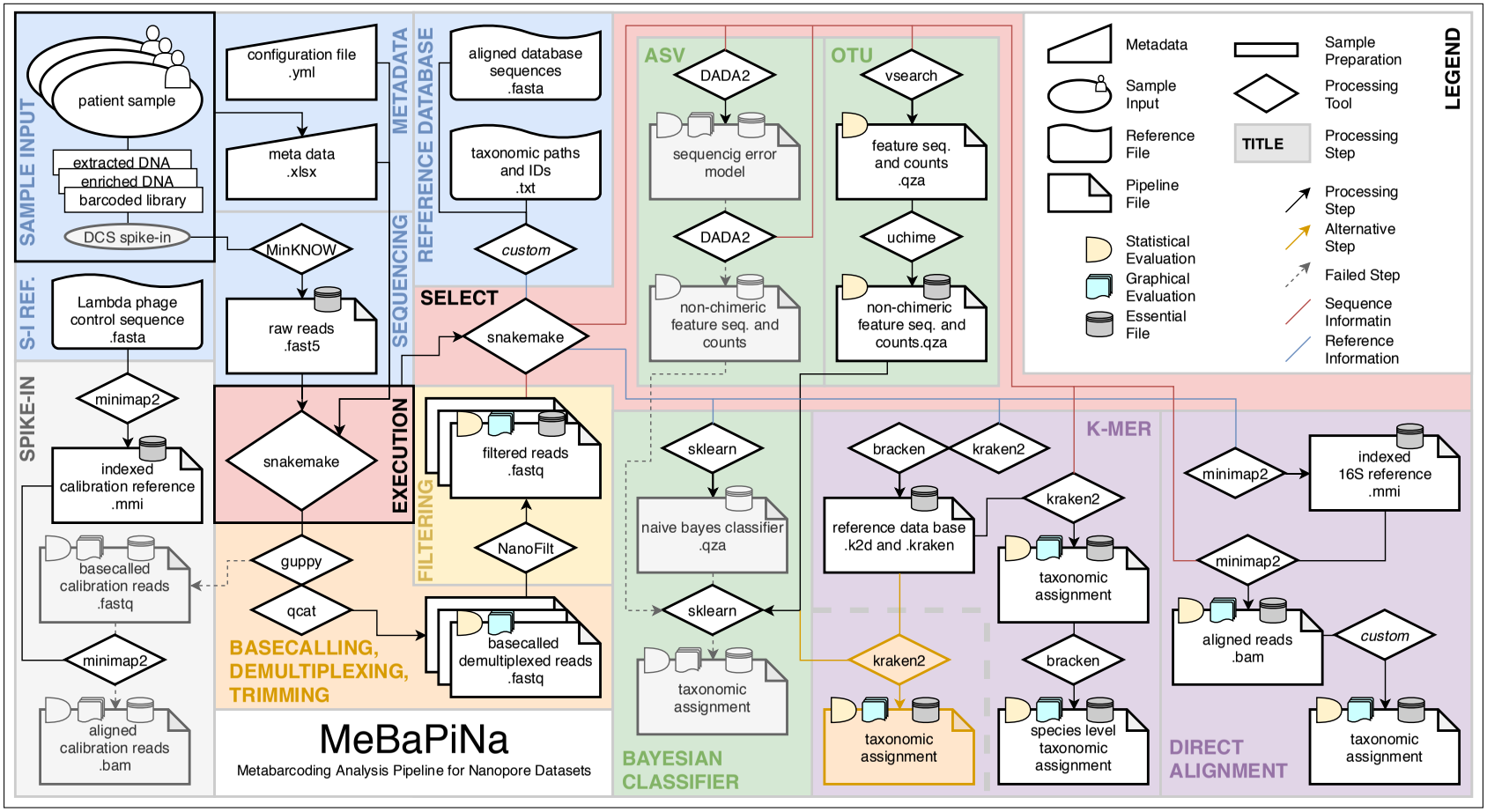

FIGURE 2. MeBaPiNa workflow: A schematic overview of the MeBaPiNa (metabarcoding analysis pipeline for Nanopore datasets) workflow. It is separated into steps indicated by the colored boxes. Blue boxes for input steps performed prior to pipeline execution, red for central pipeline steps, orange and yellow for quality control, green and purple for feature extraction and taxonomic classification methodologies.

TABLE 3. Read statistics: A list of read statistics at multiple points during the quality control. Shown are the throughput as total number of reads and total bases, the median length and the median quality after basecalling (bac), after demultiplexing (dem) and after trimming and filtering (flt). The values after demultiplexing (dem) exclude unassigned reads and the values indicated with "*" are averages of the sample median values of Table S1.

run	reads [10^6]			bases [Gb]			length [kb]			quality		
	bac	dem	flt	bac	dem	flt	bac	dem	flt	bac	dem	flt
I	11.9	10.1	8.6	18.2	15.7	12.4	1.6	1.4*	1.5	10.1	10.0*	11.8
II	8.5	7.2	3.2	8.3	7.1	4.6	0.9	1.2*	1.4	9.6	10.1*	11.6

Despite the large difference in read counts, the ratio of total bases between both runs (a ratio of 2.2 between run I and run II) was closer to the ratio of run durations. Further, with 1.59 kb, the median read length of run I was closer to the expected amplicon size of around 1.5 kb (see Section 1.2.2) than for run II with 0.87 kb. Together, this indicated a strong difference in the read length distribution of both runs.

As shown in Figure 3a run I showed a narrow peak of read with a length close to the expected amplicon size and a small fraction of reads at lower read lengths. Few reads were longer and most of them had low quality scores. In contrast, run II showed a bimodal read length distribution with two distinct peaks and approximately the same number of reads in both subsets (Figure 3b). One peak was of similar location and range as the single peak of run I representing the target amplicon. The second peak was at a length below 500 bp, which is outside the range of the expected amplicon, making it unlikely that the reads originated from the target region. Few reads were found in between the peaks or above the expected length, comparable to run I. The difference in the distributions of both runs resulted in the different median read length seen in Table 3.

The quality scores of both runs showed a continuous distribution with a median score of 10.1 for run I and 9.6 for run II (see Figure 3 and Table 3). This corresponds to an average base call accuracy of 90.2% and 89.0%, respectively. Besides the main distribution, both runs had a minor subset of low quality reads, below a Phred quality score of 7. The origin of this separated subset was unclear, but might be caused by artifacts created arisen during the PCR amplification of the target region. The analysis of the positional quality score revealed a decreased quality at the read

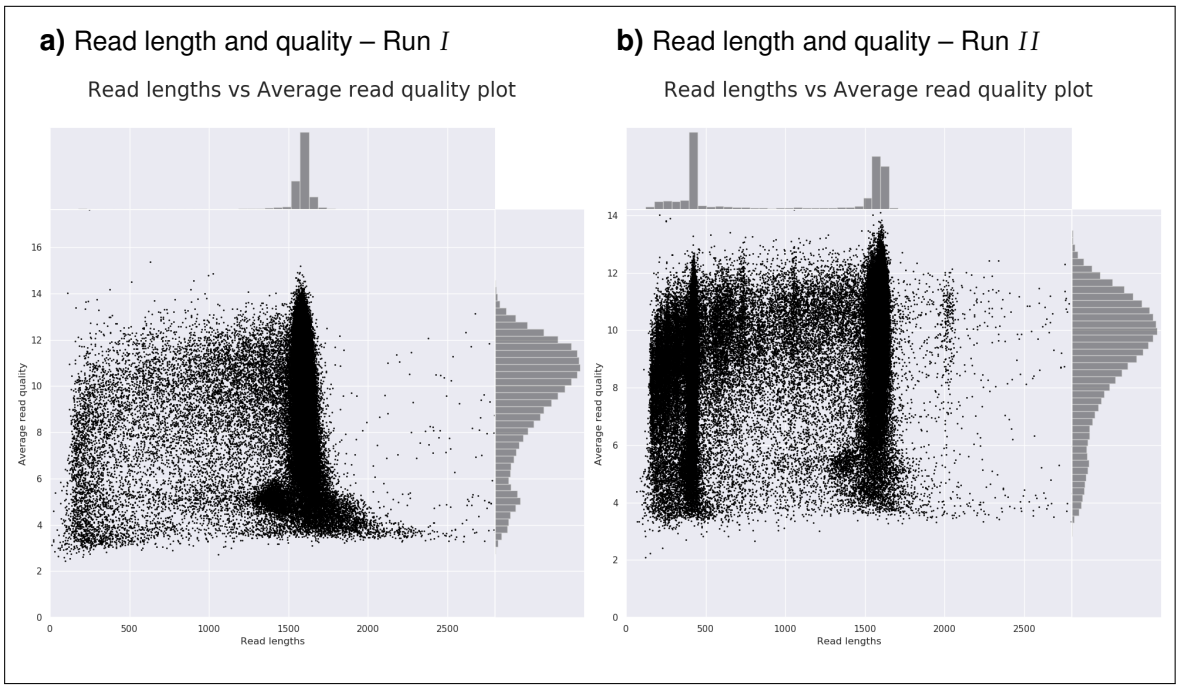

FIGURE 3. Read length and quality distribution after basecalling: A scatter plot of the average read quality (given as Phred score) over the read length for the reads of run I (Subfigure **a)**) and run II (Subfigure **b)**). The panel above the scatter plot shows the read length distribution histogram, the panel on the right a histogram of the quality distribution. The read length was cut off at 2800 bp.

ends, especially over the first 40-50 bp of the reads (Figure S3).

Overall, the basecalling step of both runs was completed successfully, but the resulting read length and quality distribution showed the need of filtering to exclude reads with low quality or outside the expected amplicon size as well as trimming of low quality read ends before proceeding with downstream analysis.

2.3.2 Demultiplexing and Trimming

All reads were processed together during basecalling, independent of the sample they originate from. In order to separate the sample information and analyze them independently, the reads had to be demultiplexed before trimming of the read ends containing the adapters and barcodes. Therefore, the barcodes at the read ends were detected and compared to the provided metadata to assign the reads to the corresponding samples.

85% of the reads (10.1*10^6 reads for run I and 7.2*10^6 reads for run II) comprising 86% of the total bases (15.7 Gb for run I and 7.1 Gb for run II) were successfully assigned to a sample during demultiplexing (see Table 3). All barcodes were detected at least once during demultiplexing, meaning no sample dropouts occurred.

However, the reads were not evenly distributed over the samples. Between 0.6*10^6 and 1.8*10^6 reads (a total between 1.0 Gb and 2.8 Gb) were assigned to the non-control samples of run I (see Table S1 and barcode03-barcode12 in Figure 4a). The median read length and quality was comparable across all non-control samples and in agreement with the characteristics described after basecalling (compare barcode03-barcode12 in Figure 5a and 5c to Figure 3). No difference between the sample sources was observed in run I.

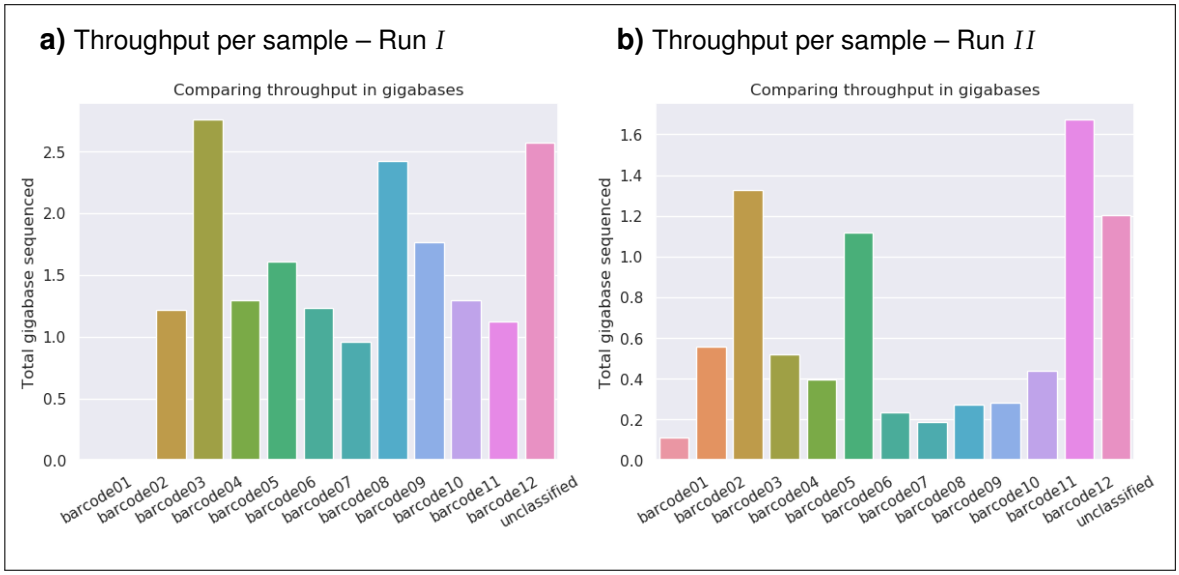

FIGURE 4. Sequencing throughput per sample after demultiplexing: Shown is the sample throughput as total bases (in Gb) after demultiplexing of run I (Subfigure **a)**) and run II (Subfigure **b)**). The color has no association and only helps visual separation of the bars. Be aware that the labels are centered underneath the bar they correspond to, which is not configurable in the tool producing the plots (see Section 3.5).

A minimal number of reads (44 for the no-template control and 30 for the extraction control) with a median length of only about 200 bp and below overall median quality of 7.5-7.7 were assigned to the controls (see Table S1 and barocde01/barcode02 in Figure 4a). Closer inspection revealed a mix of two read length subsets with mostly low quality (barocde01/barcode02 in Figure 5a and 5c), indicating a mixture

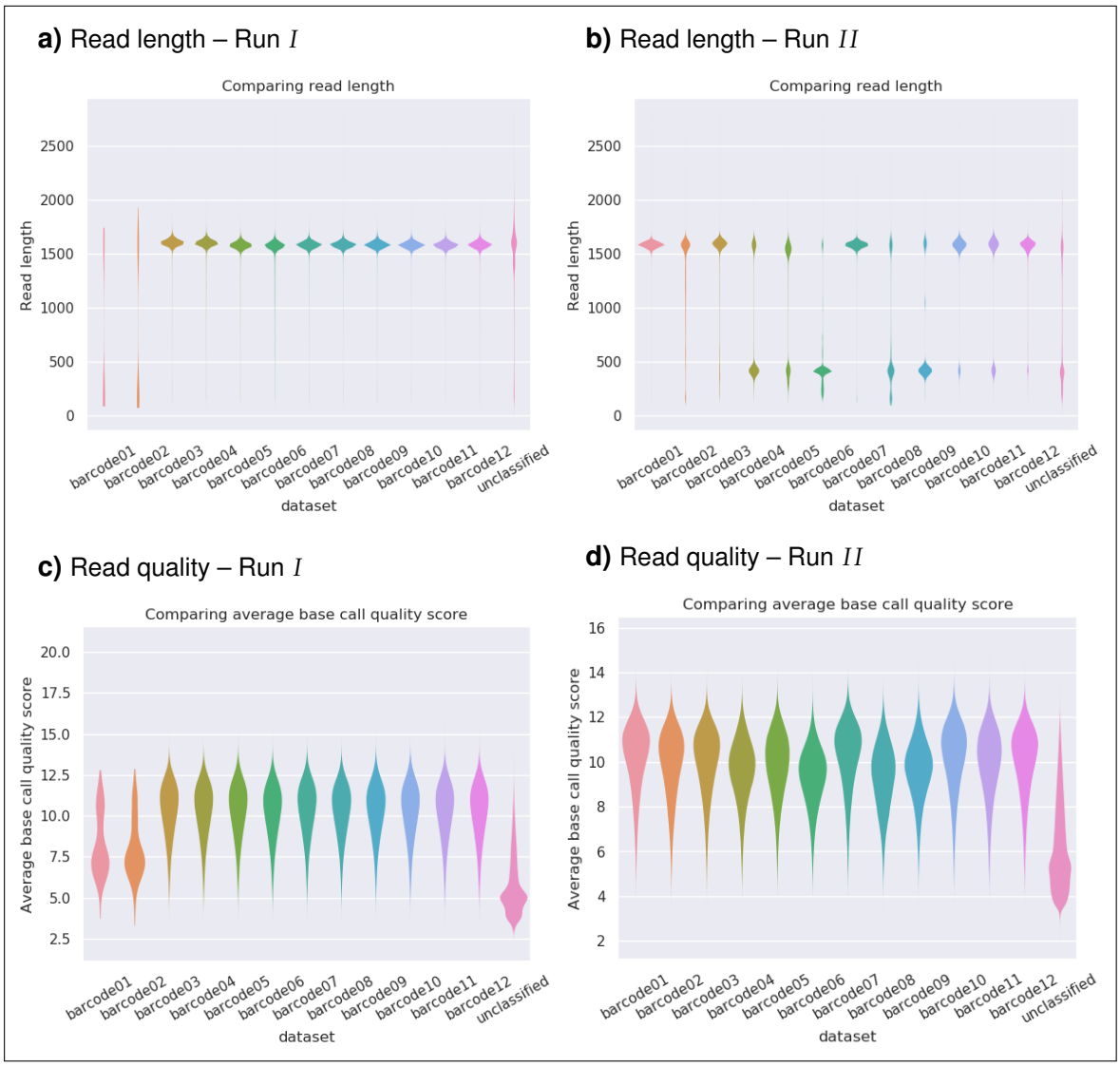

FIGURE 5. Read length and quality distribution per samples after demultiplexing: Shown are the sample read length (Subfigure **a)** and **b)**) and quality distribution (Subfigure **c)** and **d)**) after demultiplexing of run I (Subfigure **a)** and **c)**) and run II (Subfigure **b)** and **d)**). The sample-barcode-associations are shown in Table 1. "unclassified" specifies reads excluded during demultiplexing. The color has no association and only helps visual separation of the violin plots. Be aware that the labels are centered underneath the violin plot they correspond to, which is not configurable in the tool producing the plots (see Section 3.5).

of reads with the correct amplicon length and potential artifacts introduced during PCR, sequencing or through contamination.

The non-control samples of run II showed larger differences in coverage with between 0.2*10^6 an 2.2*10^6 reads (a total between 0.2 Gb and 1.7 Gb) (see

Table S1 and barcode03-barcode12 in Figure 4b). Most samples had relatively low throughput, whilst few were assigned a large number of total bases. Independent of the throughput, all tissue samples showed a mixture of two read length subsets as observed after basecalling (compare barcode04-barcode06 and barcode08-barcode12 in Figure 5b to Figure 3). As described before, the subset of short reads is outside the range of the expected amplification length of around 1.5 kb (see Section 1.2.2) and had a strong influence on the samples median read length and slightly reduced quality (see Table S1 and Figure 5b and 5d). The enrichment performed on the extracted DNA to remove host DNA contamination, which can cause undesired amplifications and reduce throughput, showed an effect for some samples. The read length distribution of Lu18, which had the largest short read subset, showed an increased fraction of the expected read length upon enrichment (compare barcode06 to barcode10 in Figure 5). The same was observed for Lu5 (compare barcode04 to barcode08 and barcode11 in Figure 5), but not for Lu13 where enrichment even increased the number of short reads (compare barcode05 to barcode09 in Figure 5). It was therefore unclear, whether the short read subset originated from host DNA contamination. No association between the enrichment and the amount of throughput was observed and the benefit of the enrichment remained uncertain. The mock community sample showed a read length and quality distribution comparable to the samples of run I, without an additional short read subset. As the short read subsets were only found in the tissue samples, differences of sample source and library preparation seemed to be responsible for the differences seen in the distributions of both runs.

A considerable number of reads was assigned to the controls during demultiplexing of run II. 0.1*10^6 and 0.4*10^6 reads (a total of 0.1 Gb and 0.6 Gb) were assigned to the not enriched and 0.2*10^6 reads (a total of 0.2 Gb) to the enriched control (see Table S1 and barcode01/barcode02 in Figure 4b). The read length and quality was within the expected range which suggested a contamination or transfer of DNA from one sample to another during the DNA extraction or library preparation or an incorrect assignment of the read barcodes (referred to as sample bleeding or sample cross-talk).

Further analysis of the unassigned reads of both runs revealed that they were mostly at a similar range as the assigned reads (see Figure 5b and 5a), but of lower quality (see Figure 5c and 5d). Hence, the reads were likely correctly amplified, but failed

to be assigned to barcodes because of their reduced quality, rather than originated from truncated reads caused by degradation of the library.

The read ends consisting of the sequencing adapter, barcode and the primer used for the amplification of the target region were artificially added to the template sequences during library preparation and could influence the downstream analysis. Therefore, both ends of the reads were trimmed up to the template sequence after the demultiplexing was completed. Trimming also removed low quality bases at the read ends detected in the basecalled reads (as seen in Section 2.3.1).

Reads from all samples were disjoined during demultiplexing and part of the reads with low quality were excluded during demultiplexing. However, the subset of reads with length below the expected amplicon range persisted and further filtering was required.

2.3.3 Length and Quality Filtering

Reads of low quality and length outside the expected range were detected in both runs and had to be removed to improve performance of downstream processes. A read length and quality filtering was performed based on a set of parameters provided as part of the configuration file. Based on the expected amplicon length, reads below 1 kb and above 2.8 kb were filtered out. Additionally, reads below a Phred quality score of 7 were also excluded.

After filtering, 8.6*10^6 reads (equal to 72% of basecalled or 85% of demultiplexed reads) were retained for run I comprising a total 12.4 Gb (equal to 68% of basecalled or 79% of demultiplexed total bases) (see Table 3). With 3.2*10^6 reads (equal to 37% of basecalled or 44% of demultiplexed reads) and 4.6 Gb (equal to 55% of basecalled or 65% of demultiplexed total bases), fewer reads were retained during filtering of run II, due to the exclusion of the subset of shorter reads (compare Figure 3 and S4). Thus, the initial read throughput ratio between run I to run II shifted from 1.4 to 2.7, closer to the ratio of run duration of 2.5. The median read length of both runs was between 1.4 kb and 1.5 kb and the median quality increased to 11.8 and 11.6 for run I and II (see Table 3 and S4), corresponding to an average base call accuracy of 93.4% and 93.1% respectively.

After filtering, the throughput for non-control samples was generally reduced slightly retaining between 0.5*10^6 and 1.5*10^6 reads (a total between 0.8 Gb and 2.2 Gb) for run I and between 0.1*10^6 and 0.9*10^6 reads (a total between 0.1 Gb and 1.3 Gb) for run II (see barcode03-barcode12 in Figure 6 and Table S1). The reduction was higher for run II as a higher fraction of short reads were removed here. The not enriched sample of Lu18 (run II, barcode 06) was an extreme example for this trend. It had a large fraction of reads with short read length which were excluded during filtering and its throughput dropped from 2.2*10^6 reads with 1.1 Gb to 0.2*10^6 reads with 0.2 Gb. The mock communities were well represented in both runs. The cell based mock community showed the highest number of retained bases in run I and the DNA based community was associated with reads equivalent to 1 Gb in both runs. No trends in the distribution of the stool samples was observed, neither was an effect of the DNA enrichment. Of note was the large fraction of reads retained for the enriched ovary sample in run II.

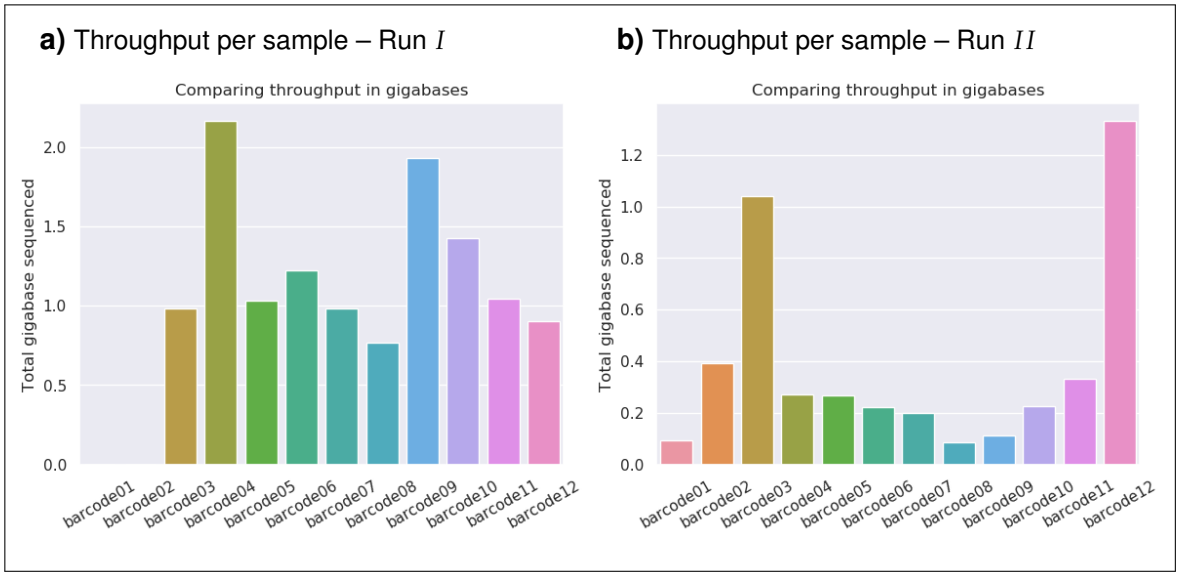

FIGURE 6. Sequencing throughput per sample after filtering: Shown is the sample throughput as total bases after filtering of run I (Subfigure **a)**) and run II (Subfigure **b)**). The color has no association and only helps visual separation of the bars. Be aware that the labels are centered underneath the bar they correspond to, which is not configurable in the tool producing the plots (see Section 3.5).

All samples of both runs showed an enrichment of reads in the expected length of the target amplicon (see Figure 7a and 7b) and similar average qualities (see Figure 7c and 7d), following the overall trend.

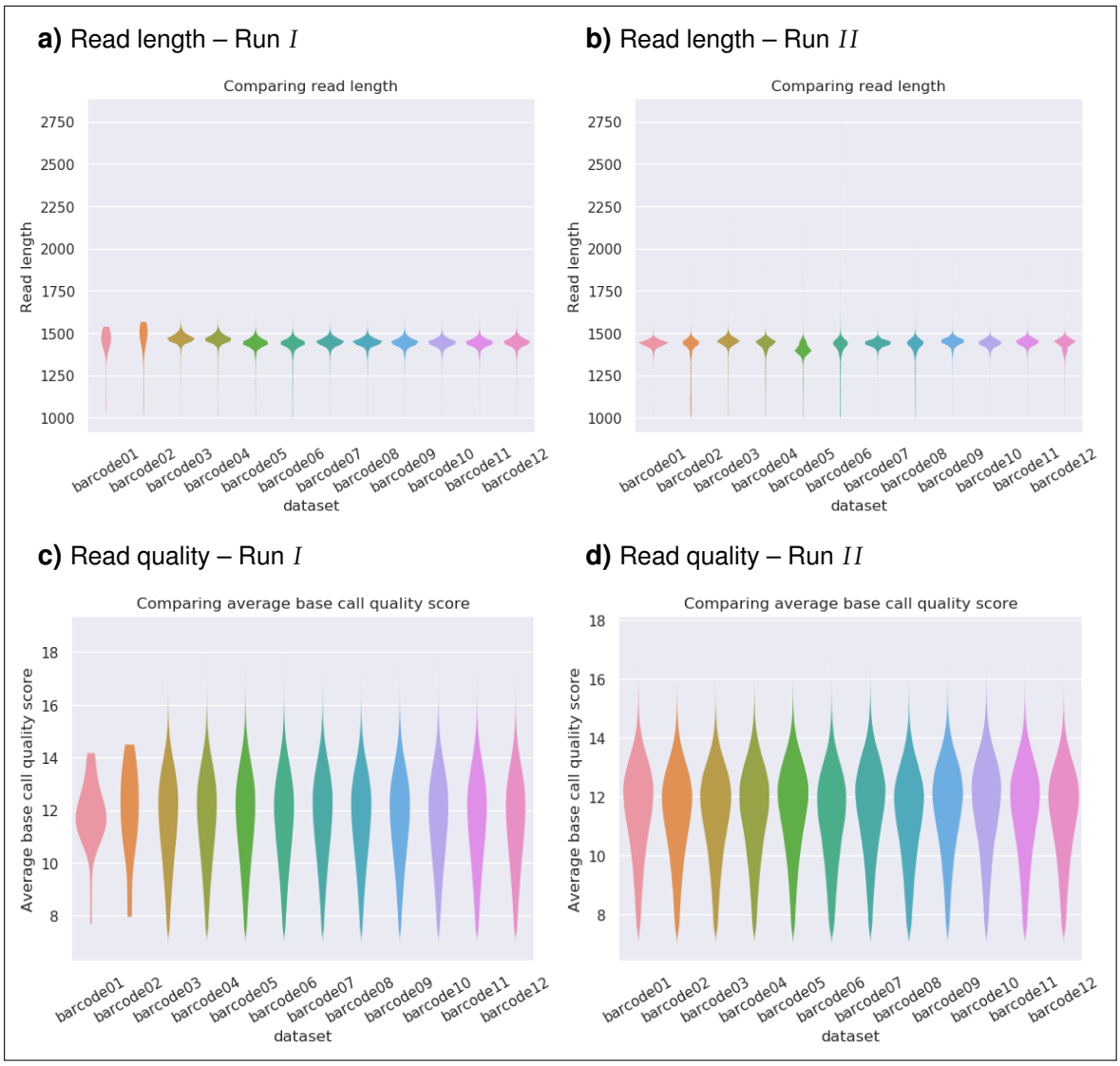

FIGURE 7. Read length and quality distribution per samples after filtering: Shown are the sample read length (Subfigure **a)** and **b)**) and quality distribution (Subfigure **c)** and **d)**) after filtering of run I (Subfigure **a)** and **c)**) and run II (Subfigure **b)** and **d)**). The sample-barcode-associations are shown in Table 1. The color has no association and only helps visual separation of the violin plots. Be aware that the labels are centered underneath the violin plot they correspond to, which is not configurable in the tool producing the plots (see Section 3.5).

After filtering, the quality control was completed. More than half of the total bases produced in both runs were retained, despite the exclusion of a large fraction of reads from run II. The remaining reads were of good quality and showed a length distribution matching the expected amplicon size.

2.4 Feature Extraction and Taxonomic Classification

In order to analyze the microbial communities in the samples, a taxonomy had to be assigned to the reads after quality control and the frequency of each taxon had to be determined. Four different methodologies were implemented in MeBaPiNa (see MeBaPiNa workflow in Figure 2). Two of which used the quality controlled reads as input directly, the other two aimed to extract features from the reads first, before classifying the taxonomy.

The method of choice is selected through the corresponding parameter in the configuration file and the pipeline automatically executes the rules necessary for the feature extraction and taxonomic classification build around the selected methodology.

2.4.1 Reference Database

Independent of the methodology, the inference of taxonomy required a set of sequences with know taxonomic assignment as reference. The sequence information from the samples was compared to this reference dataset to find a appropriate taxonomic classification.

A modified version of the most recent SILVA SSU database release was used as a reference dataset in MeBaPiNa. The required reference files were automatically downloaded and processed. A combination of tools and custom scripts was used to reduce all reference sequences to the amplicon region, filter them by length and remove duplicates. The taxonomic assignments were extended to species rank, if necessary. A separate instance of the reference database without species level association was created in parallel and could be selected as reference in the configuration file.

The SILVA SSU database consisted of 510'984 reference sequences. During filtering, 59'138 were excluded leaving a total of 451'846 unique sequences associated with 10'258 higher taxa. The addition of species level taxonomy increased this initial number to 90'883 taxa of which 78'191 were unique species assignments and 12'692 were unique higher level assignments listed in Table 4.

TABLE 4. Reference taxonomic ranks: List of taxonomic ranks represented in the modified reference database, their depth from the root and the number of unique taxa at this rank. The rank "other" refers to higher taxa without specific rank or taxa in between the specified ranks.

depth	rank	count
0	root	1
1	domain	3
2	phylum	220
3	class	523
4	order	1'183
5	family	1'429
6	genus	6'679
7	species	78'191
-	other	2'654

Based on this combination of reference sequences and their taxonomic assignment, method specific database formats were created and used to identify features in the sequencing reads and classify them.

2.4.2 K-mer Mapping

Taxonomic Classification with Kraken 2

The k-mer mapping approach is one of the methodologies used for direct taxonomic classification of the quality controlled sequences. Each read was used as an individual input feature for the Kraken 2 utility, which split them into k-mers and compared them to a database trained on the SILVA reference sequences. The reads were classified based on the reference match with lowest confident taxonomic rank.

The processing with eight parallel threads took a maximum 5:26 min per sample and between 97 an 100% all of the reads were successfully classified (compare Table 5 and S1). The majority of reads of both runs was assigned to species taxa and only a small fraction of reads was assigned to higher taxa under the chosen parameters, indicating a good resolution of this approach.

For the non-control samples of run I, between 1'155 and 3'476 taxa were classified of which 89%-95% were of species rank (see Table 5). However, as no further filtering was included in this step, a high number of taxa were singletons (only one

TABLE 5. Taxonomic coverage – k-mer mapping: The number of classified reads and taxa as well as the median reads per taxa for all assigned taxa (assign.), species rank taxa (spec.) and species after reestimation of abundance with Bracken (reest.) of each sample (run I top, run II bottom, see Table 1 for reference).

ID	reads [10^6]			taxa			median	
	assign.	spec.	reest.	assign.	spec.	reest.	spec.	reest.
NTC	<0.1	<0.1	<0.1	28	22	3	1	4
EC	<0.1	<0.1	<0.1	19	15	3	1	4
ZyDNA	0.7	0.6	0.7	1’155	1’100	637	4	16
ZyCell	1.5	1.4	1.5	1’540	1’454	863	4	18
St61	0.7	0.6	0.7	2’453	2’216	1’239	3	16
St01-1	0.9	0.8	0.8	2’984	2’665	1’492	3	15
St02-1	0.7	0.6	0.7	2’440	2’182	1’178	3	14
St01-2	0.5	0.5	0.5	2’407	2’149	1’145	3	13
St03-1	1.3	1.2	1.3	3’476	3’141	1’712	3	14
St04-1	1.0	0.9	1.0	2’595	2’320	1’302	3	13
St05-1	0.7	0.7	0.7	2’455	2’199	1’197	3	14
St02-2	0.6	0.6	0.6	2’468	2’217	1’216	3	14
NTC	0.1	0.1	0.1	709	663	311	2	10
EC	0.3	0.3	0.3	1’472	1’370	712	3	13
ZyDNA	0.7	0.7	0.7	1’406	1’326	750	3	16
Lu05	0.2	0.2	0.2	1’431	1’337	636	2	10
Lu13	0.2	0.2	0.2	1’958	1’793	874	2	11
Lu18	0.2	0.1	0.1	1’993	1’828	796	2	10
EC-en	0.1	0.1	0.1	922	853	422	2	12
Lu05-en	0.1	0.1	0.1	885	823	376	2	11
Lu13-en	0.1	0.1	0.1	644	585	236	2	13
Lu18-en	0.2	0.1	0.2	701	638	288	2	8
Lu05-en2	0.2	0.2	0.2	1’583	1’457	756	3	13
Ov85-en	0.9	0.9	0.9	2’540	2’402	1’363	3	13

read was assigned to them) reflected in a low median number of assigned reads per species. Only minor differences in the number of classified taxa per clinical samples was visible and the number of taxa seemed to be exclusively dependent on the number of assigned reads. At comparable assigned read counts, fewer taxa were classified for the mock community compared to the clinical samples of run I, resembling the lower complexity of the mock community. However, the number of taxa was far above the number of species comprising the community. The small number of, mostly singleton, taxa detected in the control samples could not explain this inflation.

The non-control samples of run II had between 644 and 2'540 classified taxa of which 91%-95% were species rank taxa (see Table 5). Similar to run I, a high number of singletons was observed resulting in a low number of median reads per species taxon. A large quantitative difference between the samples of enrichment batch I and their native counterparts was observed, showing a reduction in complexity after enrichment. However, the same difference was observed in the number of reads assigned to the taxa and no difference in median read assignment was visible. Normalizing the number of classified taxa to the assigned reads, increases the difference between the clinical samples and mock community, following the trend of lower complexity in the mock community, observed in run I. The ovary sample remained the sample with the highest representation in run II, reflecting its larger abundance in the quality controlled reads. As initially observed during demultiplexing and filtering (see Section 2.3.2 and 2.3.3), the control samples of run II showed higher read counts which propagated into a high number of classified taxa in both control samples and suggested the presence of a kit contamination in the sequencing library.

The classification via k-mer mapping resulted in a high number of assigned reads and high taxonomic resolution. However, a large number of singletons was observed.

Reestimation of Abundance

The taxonomic classification via k-mer mapping was further optimized by a Bayesian reestimation of abundance implemented in Bracken. The reestimation was based on the initial classification results of the k-mer mapping and the sequence information in the reference database. Reads of all taxa were reclassified to species rank and low abundance taxa were filtered.

The reestimation had a strong influence on the number of classified taxa, but not on the number of assigned reads (see Table 5). About half of the classified taxa of all non-control samples were reassigned or filtered out retainnig between 40% and 59% of all classified taxa. Nevertheless, the trends observed for the initial classification remained the same with the number of taxa following the number of assigned reads and the mock communities showing lower complexity. However, despite the reduction, the number of classified taxa for the mock community samples remained

high (637-863 species) compared to its true composition fo eight individual species. The number of taxa in the controls of run I were further reduced, but the controls of run II remained comparable to the clinical samples. The overall change to the taxa composition was reflected by an increase in the median read assignment per taxon. After filtering, the clinical samples of run I had a higher median read count per taxa then the clinical samples of run II. Together this suggests that an increase in read counts leads to an increase in classification of taxa with low read counts, partially due to the accumulation of erroneous reads, but also due to the increased detection of low abundance species.

Reestimation and filtering successfully reduced the number of taxa with low abundance and reassigned reads from higher ranks to species taxa resulting in a total reduction of taxa of up to 60%, whilst requiring only seconds per sample to complete.

Mock Community

To analyze the effect of the k-mer mapping approach, the reestimated classification of the mock community samples was further analyzed and compared to the reference composition.

All reference species taxa were detected in the mock community samples of both runs, as shown in Figure 8a. However, the fraction of reads assigned to the correct reference taxa was low for all samples and 74%-79% were assigned to other taxa. Consequently, the relative deviation from the input fraction showed a depletion of all taxa (see Figure 8b). Due to its small proportion in the input reference, *Pseudomonas aeruginosa* was the spices with the strongest relative deviation in all three samples. In contrast, *Listeria monocytogenes* was assigned to between 9%-7% of all reads, corresponding to a 2.1-1.6 fold relative depletion. No clear difference between the runs or sample sources was observed.

The investigation of the full taxonomic classification (shown exemplarily for ZyDNA of run I in Figure 9 and for the other samples in S5 and S6), revealed that large numbers of reads were assigned to species other than the reference taxa. Despite the large number of taxa observed during the classification, few species were detected in high numbers, as expected from the reference community composition. Some of the reads were classified as taxa closely related to the input species, e.g.

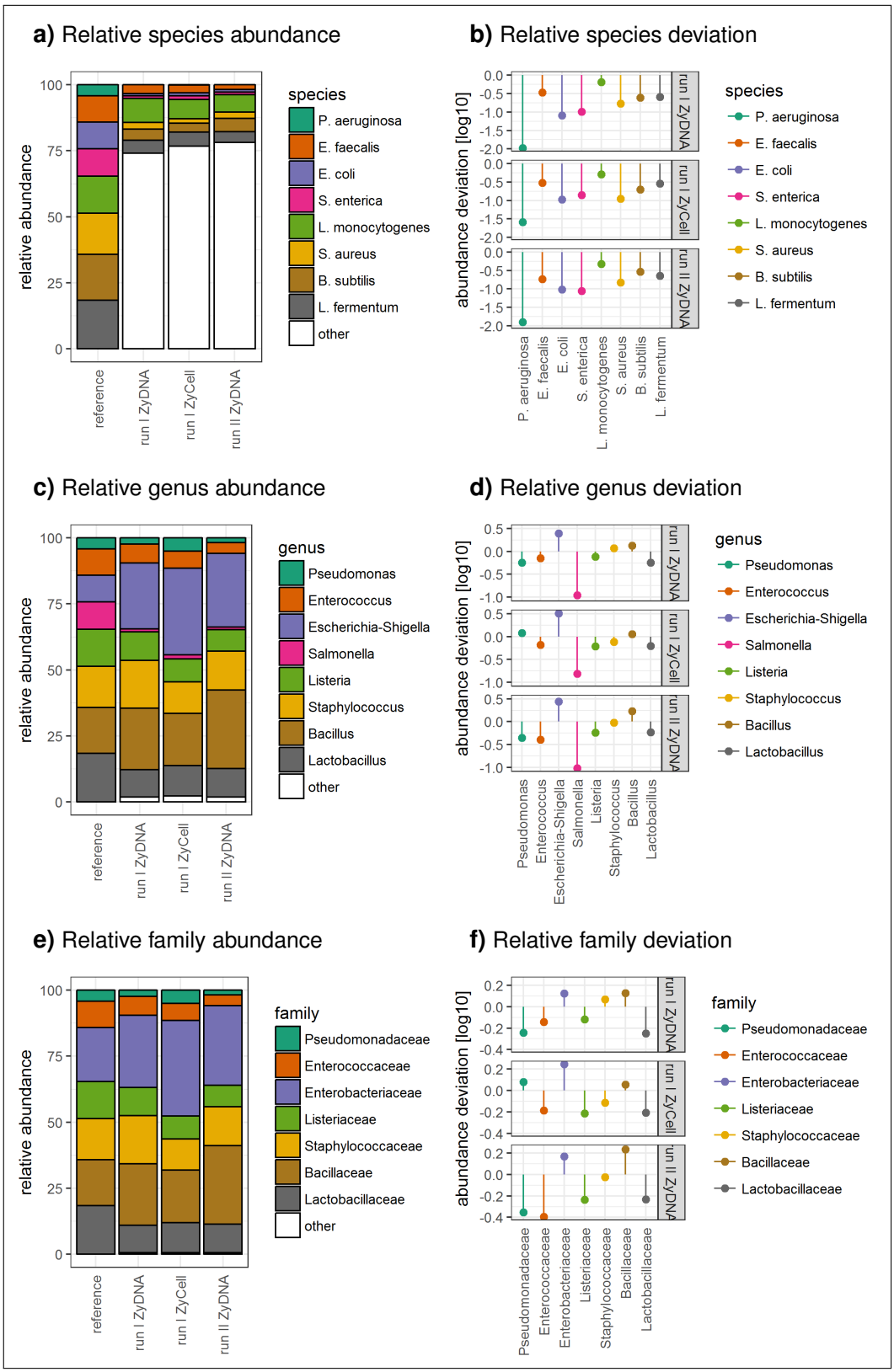

FIGURE 8. Mock community taxonomic abundance and deviation – k-mer mapping: Shown is the relative taxonomic abundance of the reference taxa (Subfigure **a)**, **c)** and **e)**) for the three mock community samples and the reference as well as their relative deviation (Subfigure **b)**, **d)** and **f)**) for species (Subfigure **a)** and **b)**), genus (Subfigure **c)** and **d)**) and family (Subfigure **e)** and **f)**) ranks.

Enterococcus faecium instead of *Enterococcus faecalis* or *Listeria innocua* instead of *Listeria monocytogenes*, accounting for some of the deviation seen for the species rank. However, many of the classified taxa with high abundance were of uncertain origin (e.g. "metagenome" or "uncultured bacterium"), but showed sizes comparable to the expected reference taxon of the same genus, which indicates a incorrect assignment of the reads to these ambiguous taxa.

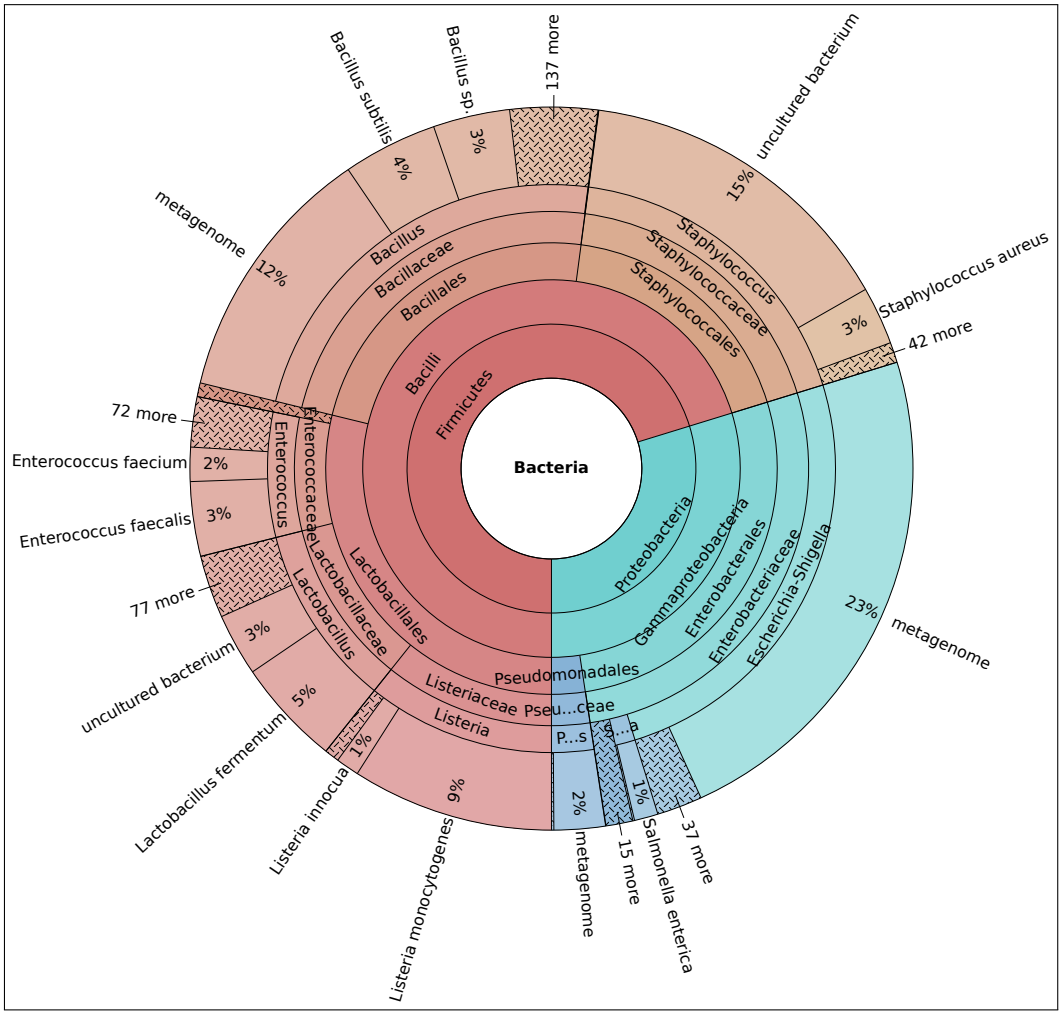

FIGURE 9. Mock community taxonomic composition – k-mer mapping: Multi-layer pie chart of the taxonomic composition for the mock community, exemplarily shown for ZyDNA of run I using k-mer mapping and reestimation of abundance. Slices indicate the relative abundance of the taxa at each rank (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards from higher taxa represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

This was further confirmed by the relative taxonomic abundance at genus rank (see

Figure 8c). Here, the fraction of reads not assigned to reference taxa was at 2% for all samples. Except for *Escherichia-Shigella* and *Salmonella*, most genera showed low deviation from the reference in all samples (see Figure 8d). *Escherichia-Shigella* showed 2.5-2.8 times the number of expected reads in the ZyDNA samples and even 3.2 times in the ZyCell sample. In contrast *Salmonella* showed a 10 fold decrease in the DNA mock community samples and a 6.7 fold decrease in the cell based sample. As both genera shifted in opposite directions and are closely related, an assignment of reads from one genus to the other is likely. Besides the distortion of these genera, the ZyCell sample showed overall lower deviations of the genera. This could be a result of the different sample source or the higher read counts of this sample (see Section 2.3.3).

The assignment to family taxa confirmed the results of the genera classification (see Figure 8d and Figure 8e). The fraction of reads assigned to families other than the target ones was below 1% and the strong deviations seen for the genera were canceled out as both *Escherichia-Shigella* and *Salmonella* belong to the *Enterococcaceae* family. Generally, the deviations for family taxa were lower than for the genera and lower for the samples of run I than of run II, possibly due to the potential kit contamination in run II.

The samples also showed close relation in their statistical measurements (see Table 6). The species richness, reported as the number of taxa, was similar between the samples, but vastly higher than the for reference. The diversity measures were nearly identical for all samples and indicated a low diversity, which closely reflects the original composition of the community, as a similar number of highly abundant, but incorrectly classified, taxa were detected. The mixture of low abundance and high abundance taxa was reflected in a medial evenness in the mock community samples and showed a strong reduction compared to the reference.

TABLE 6. Mock community properties – k-mer mapping: Listed are the species richness (S), the diversity as Shannon entropy (H) and Simpson index (λ) as well as Pielou's evenness (J) in all samples and the underlying community reference.

run	ID	S	H	λ	J
I	ZyDNA	637	2.99	0.89	0.46
I	ZyCell	863	2.96	0.88	0.44
II	ZyDNA	750	3.05	0.89	0.46
-	reference	8	2.01	0.86	0.97

Despite the high number of assigned taxa, the assignment showed a good accordance to the reference composition for the genus and family ranks and similar overall properties. However, large fractions of the reads to be assigned to undescriptive species taxa, reducing the information content at this rank.

2.4.3 Full-Length Alignment

Alignment with minimap2

The alignment approach is similar to the k-mer mapping as it processed the reads as individual features and assigned taxonomy based on sequence similarity to the SILVA reference. However, instead of splitting the reads into k-mers a full-length read alignment was performed with minimap2. The primary alignment of the reads with a unique highest alignment score were kept and assigned to the taxonomy of the reference they aligned to.

The alignment was successful for all samples, but required up to 1:48 h of processing time with eight parallel threads. Between 85% and 98% of the reads per sample were aligned with a unique highest score and retained for taxonomic classification (compare Table 7 and S1). The reads showed combined alignment error rates below 10%, with the exception of the mock community sample of run II showing an error rate of 11.2% (see Table S2), caused by an increased number of mismatches. Other than that, the error rates were distributed evenly between insertions, deletions and mismatches with a tendency towards mismatch errors, which could have originated from sequencing error or misalignment. Run II showed slightly less insertions than run I, but slightly higher rates of the other error types.

The median alignment identity matched the combined error rates and were over 89% for all samples (see Table S2). Compared to the expected identities, calculated from the Phred quality scores of the filtered reads (see Table S1), the observed identities were slightly lower. As the Phred score of the Nanopore reads is only an estimation of the real identities, it remained unclear whether this was an overestimation or if the decreased identity was due to sub-optimal alignments. When comparing the distribution of Phred quality scores to the alignment identity as shown in Figure 10, the distribution followed the expected logarithmic course.

TABLE 7. Taxonomic coverage – full-length alignment: The number of successfully aligned (align.) assigned (assign.) reads, the number of classified taxa as well as the median reads per taxa of each sample (run I top, run II bottom, see Table 1 for reference).

ID	reads [10^6]		taxa	median
	align.	assign.		
NTC	<0.1	<0.1	3	3
EC	<0.1	-	-	-
ZyDNA	0.6	0.5	120	37
ZyCell	1.4	1.2	171	44
St61	0.7	0.6	447	60
St01-1	0.8	0.7	526	72
St02-1	0.7	0.6	374	37
St01-2	0.5	0.4	298	46
St03-1	1.3	1.1	575	118
St04-1	1.0	0.8	445	60
St05-1	0.7	0.6	384	76
St02-2	0.6	0.5	336	40
NTC	0.1	0.1	43	84
EC	0.3	0.2	160	13
ZyDNA	0.7	0.5	236	20
Lu05	0.2	0.1	184	16
Lu13	0.2	0.2	103	34
Lu18	0.1	0.1	150	12
EC-en	0.1	0.1	36	224
Lu05-en	0.1	<0.1	79	26
Lu13-en	0.1	0.1	61	43
Lu18-en	0.2	0.1	39	38
Lu05-en2	0.2	0.2	22	71
Ov85-en	0.9	0.8	353	28

The alignment was successful in assigning the reads to reference sequences and gave insights into the error profiles of the samples. However, the alignment had to be translated into an taxonomic classification.

Custom Filtering and Abundance Estimation

The obtained alignments were an one-to-one association of reads and reference sequence, which had to be collapsed for each taxon in order to get an abundance estimation. Further, the reads had to be filtered to exclude short or error-rich

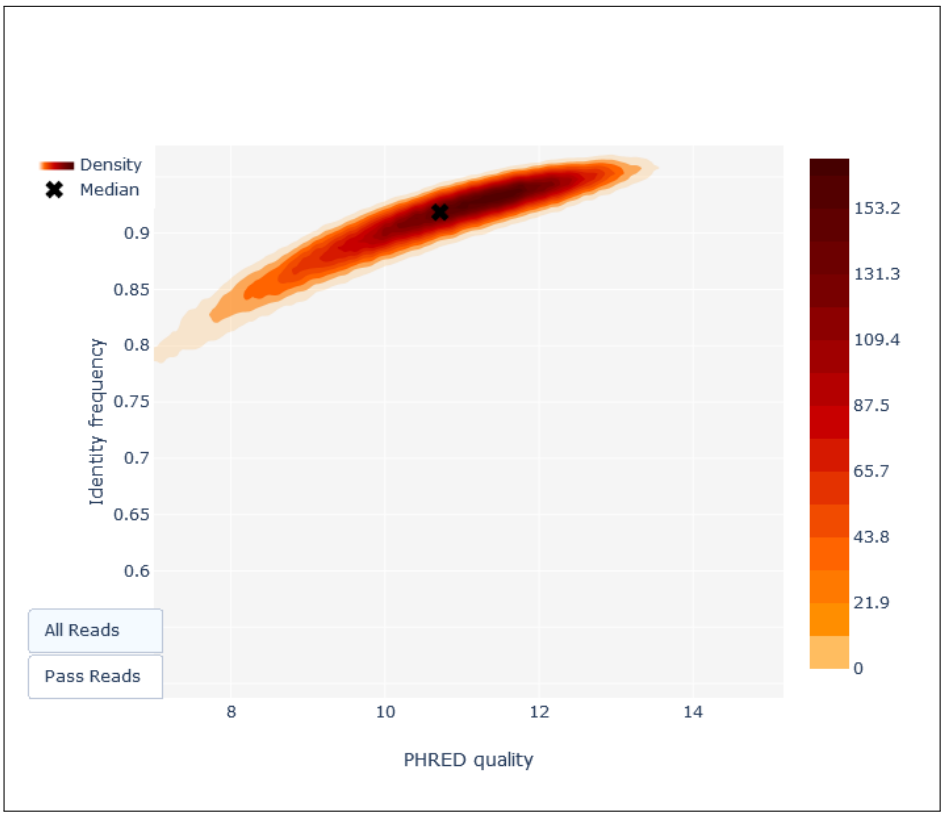

FIGURE 10. Expected and observed alignment identity: Density plot of the Phred quality score, reported by the basecaller, and the alignment identity. The "x" marks the median of both values. The Phred score Q corresponds to the error probability P with $Q = 10^{-P/10}$.

alignments. Both steps were implemented in a custom script filtering reads by length and quality and collapsing filtered reads per reference sequence and reference sequences per taxon.

After filtering, between 81% and 84% of the quality controlled reads were retained for the non-control samples of run I, corresponding to 298-575 taxa for the stool samples and 120 or 171 taxa for the mock community samples (see Table 7). During filtering, the remaining reads of the extraction control were excluded and the no-template control was reduced to a low number of taxa with few reads, unlikely to represent a contamination. The reduced richness in the mock community samples was similar to the trends seen for k-mer mapping (see Section 2.4.2). At a high percentage of assigned reads, the lower number of classified taxa was also reflected by a higher median read count per taxa (between 3 and 224), compared to k-mer mapping (between 4 and 18).

The non-control samples of run II showed higher deviations. Between 65% and 86% of the quality controlled reads were retained and between 22 and 353 taxa were classified (see Table 7). In contrast to the observations for the k-mer mapping in Section 2.4.2, the differences in classified taxa in run II could not be explained by the number of assigned reads. At comparable read numbers, the samples showed lower taxa counts and higher median read counts per taxon before than after the sample enrichment, indicating a lower complexity and higher coverage of fewer samples, similar to the mock community samples. The control samples showed a high number of classified taxa, with high median read counts, indicating the presence of a small set of taxa with relatively high coverage. Similar to run I, the mock community showed lower complexity than the not enriched clinical samples at comparable read counts. However, the number of classified taxa was higher than in run I. All mock community samples had vastly higher numbers of classified taxa than members in the input material, despite being closer to the real number than the k-mer mapping.

Collapsing the aligned reads resulted in an overall high number of classified taxa and required a maximum of 8:14 min. Clear differences were observed between the samples especially in the context of richness of the mock communities and DNA enrichment as well as between the contamination controls of both runs.

Mock Community

The mock community samples classified by the full-length alignment were analyzed to evaluate the methodology and compare it to the k-mer mapping, described in Section 2.4.2.

Between 78% and 65% of the reads were assigned to the correct reference species taxa after full-length alignment (see Figure 11a). This is in strong contrast to the low fraction of correctly assigned reads after k-mer mapping, despite using the same underlying reference database and quality controlled reads. In consequence, the observed deviations were lower (see Figure 11b). The high deviation were observed for *Pseudomonas aeruginosa*, the species with the lowest fraction in the input community, in both methodologies. The full-length alignment showed an depletion between 1.6 fold for the ZyCell sample and 10.0 fold for the ZyDNA sample of run II

for this species. On the contrary, *Salmonella enterica*, depleted in k-mer mapping, showed an overrepresentation of 1.2-1.7 fold in the full-length alignment.

The samples did not only differ from the reference and the other methodology, but also showed differences between each other. Counterintuitively, the cell based mock community showed lower deviations from the reference then the DNA based mock communities at species level, despite its more complex preparation process. Both of the DNA based communities showed the same tendencies in deviation from the reference, but the sample from run II was less accurate than the sample of run I. The lower overlap with the reference reflects the higher mismatch error rate, higher number of assigned taxa and lower median read counts per taxa in this sample, suggesting a higher number of misclassified taxa.

A detailed analysis of the samples (shown exemplarily for ZyDNA of run I in Figure 12 and for the other samples in S7 and S8) showed a representation of all reference species taxa. However, some assignment to other taxa was observed, similar to the k-mer mapping approach (see Figure 9), which explained the underrepresentation of some of the reference taxa and increased taxa counts. However, few unspecific taxa were detected, mostly of the *Escherichia-Shigella* and *Pseudomonas* genus. Other misclassified organisms included *Shigella dysenteria* and species of the *Bacillus* genus. The high fraction of *Salmonella enterica* could be linked to the underrepresentation of *Escherichia coli* in a reverse constellation than seen for the k-mer mapping. Despite the difference seen between the runs, the overall distribution of ZyDNA of run II was comparable to the samples of run I, but showed an increased number of low abundance species over all genera.

The fraction of reads assigned to the reference at genus rank was close to 99% for all samples (see Figure 11c). With an 2.6-1.8 fold underrepresentation, the largest deviation was observed for the *Lactobacillus* genus in all samples and for *Pseudomonas* in the ZyDNA samples (1.7 and 2.0 fold), but not the ZyCell sample (see Figure 11d). Both *Escherichia-Shigella* and *Salmonella* were overrepresented in the samples (1.3-1.9 fold), despite the underrepresentation of *Escherichia coli*, because of the additional species rank taxa. ZyDNA of run II showed slightly higher deviations than the samples of run I, similar to the observations for species rank associations.

This resulted in an overall good representation of the reference community at family

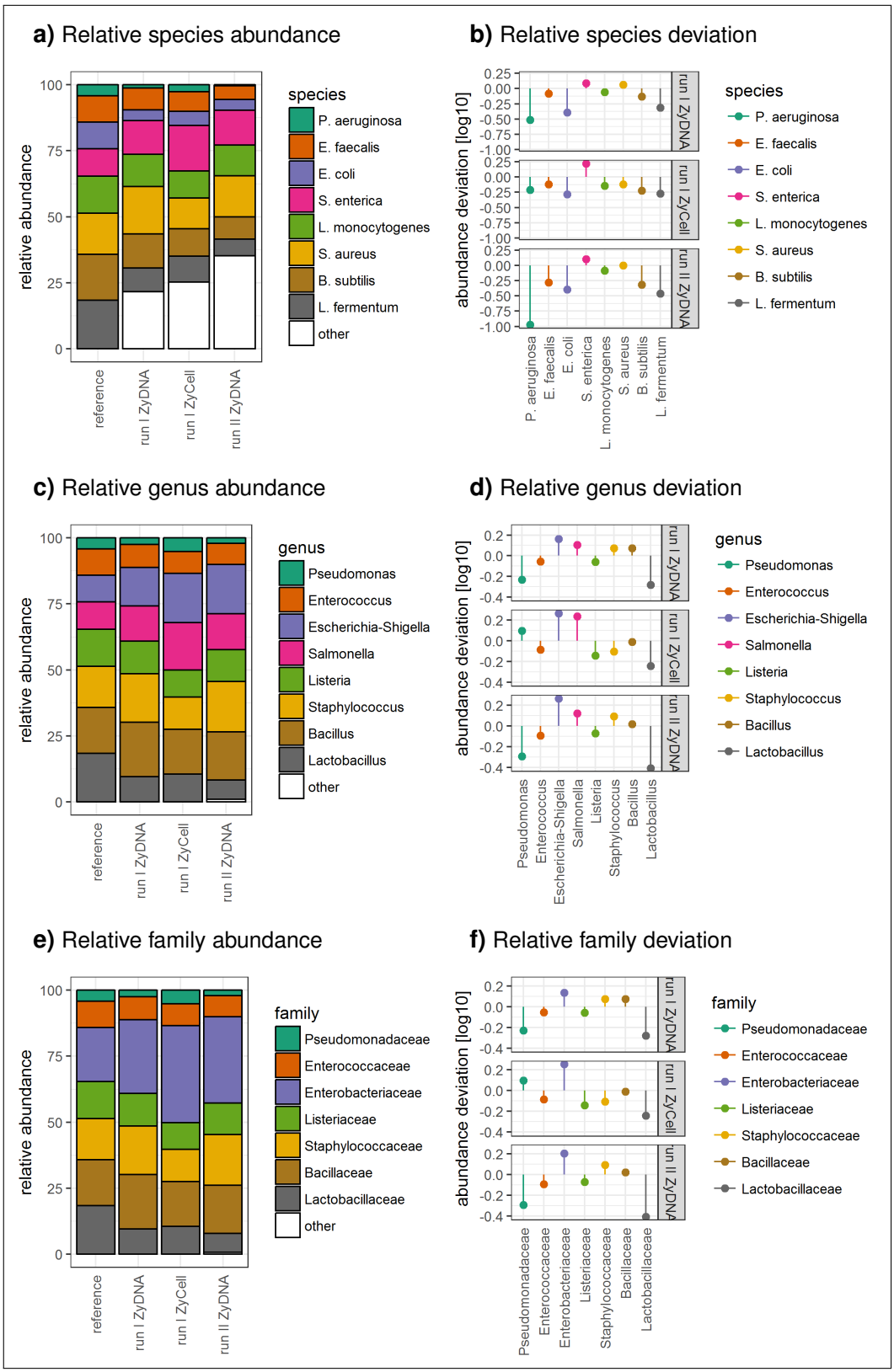

FIGURE 11. Mock community taxonomic abundance and deviation – alignment: Shown are the abundance (Subfigure **a)**, **c)** and **e)**) of the three mock community samples relative to the abundance in the input material (reference) as well as the relative deviation from it (Subfigure **b)**, **d)** and **f)**) for species (Subfigure **a)** and **b)**), genus (Subfigure **c)** and **d)**) and family (Subfigure **e)** and **f)**) ranks.

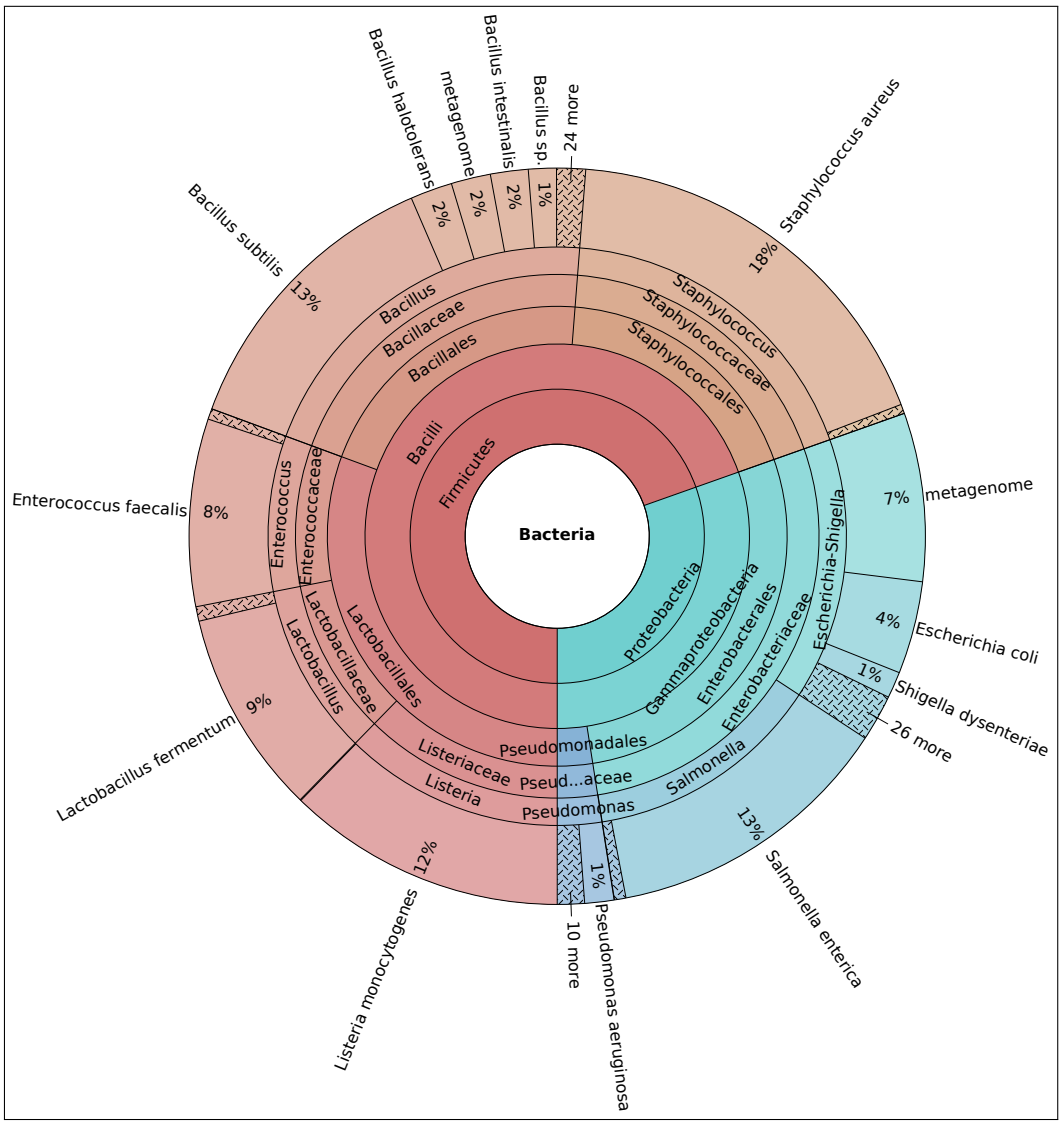

FIGURE 12. Mock community taxonomic composition – alignment: Multi-layer pie chart of the taxonomic composition for the mock community, exemplarily shown for ZyDNA of run I using full-length alignment. Slices indicate the relative abundance of the taxa at each rank (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards from higher taxa represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

rank (see Figure 11e and Figure 11f) with a higher fraction of reads assigned to members of the *Enterobacteriaceae* than expected and underrepresentations carried over from the genera distribution.

As discussed before, the species richness was lower in the alignment based classification, compared to the k-mer based approach, but still above the reference

(see Table 8). The diversity measures were comparable to results of the k-mer mapping classification and showed measures slightly above the reference. The evenness was closer to the reference for the alignment approach, as the number of low abundance taxa was reduced. The addition of multiple low abundant species in the ZyDNA sample of run II was reflected by an increase of the Shannon entropy, compared to the other samples.

TABLE 8. Mock community properties – alignment: Listed are the species richness (S), the diversity as Shannon entropy (H) and Simpson index (λ) as well as Pielou's evenness (J) in all samples and the underlying community reference.

run	ID	S	H	λ	J
I	ZyDNA	120	2.60	0.90	0.54
I	ZyCell	171	2.71	0.91	0.53
II	ZyDNA	236	2.93	0.92	0.54
-	reference	8	2.01	0.86	0.97

The full-length alignment resulted in a good representation of the reference community. In comparison to the k-mer mapping approach (up to 25.9% correctly assigned reads), a vastly higher resolution of species taxa was achieved (up to 78.4% correctly assigned reads) and overall low deviations were visible. Slight differences were observed between the runs and sample sources.

2.4.4 Operation Taxonomic Unit Picking

Feature Extraction

In contrast to the k-mer mapping and full-length alignment pipelines described in Section 2.4.2 and 2.4.3, the operational taxonomic unit (OTU) picking was performed on the raw reads to extract features and improve the subsequent taxonomic classification. The QIIME2 tool set was used to combine reads with the same sequence (dereplication) in to a single read. Then cluster dereplicated reads with a sequence identity to one of the reference sequences above 85% and use the same threshold to cluster the remaining reads by their similarity to each other (referred to as open-reference clustering). Finally, the extracted clusters were checked for chimeric sequences which were filtered out.

An initial dereplication step was performed to reduce the complexity of the quality controled read set, requiring a processing time of up to 1:05 h (parallelization was not implemented). However, due to the long read length of over 1 kb (see Table S1) and error rates of about 10% (see Table S2), all reads were unique and dereplication had no effect.

The high error rates were also the reason for the low identity threshold of 85% used during the subsequent OTU picking. The OTU picking was performed as open-reference clustering and required processing times of up to 9:17 h on 16 parallel threads. The open-reference clustering against the SILVA reference extracted a high number of up to 386*10^3 features including the entire set of input reads (see Table S3). Only a small fraction of these reads were based on a sequence of the reference and most features were de-novo clusters of input reads (up to 99.4% of the clusters per sample). Further, a large number of the features were "clusters" of single reads, resulting in a median number of reads per feature of one.

The initial set of features was, therefore, further processed to exclude chimeric sequences and filter features with a low number of assigned sequences. This step was perdormed on a single thread (parallelization was not implemented) and took up to 14:12 h to complete.

The result was a reduced set of features, listed in Table 9, usable for taxonomic classification instead of the quality controlled sequencing reads themselves. The non-control samples of run I clustered into 4.0*10^3-10.5*10^3 features containing between 68% and 75% of the quality controlled reads. All reads of the control samples of run I were excluded during this step. The trend of lower complexity in the mock communities, observed in the number of classified taxa of the previous methodologies (see Section 2.4.2 and 2.4.3), was already visible in the number of features in run I. No differences in the clinical samples were observed in run I, nor a difference in the median number of reads per feature.

The reads of run II were clustered into 0.7*10^3-7.3*10^3 features representing between 60% and 78% of the input reads. In contrast to run I, the mock community of run II showed the highest number of clustered features, besides it presumably lower complexity. At comparable number of reads, the enrichment samples of batch I showed lower complexity but comparable median read counts per feature.

TABLE 9. Taxonomic coverage – OTU picking: The number of reads per filtered feature (feature), assigned to all taxa (assign.) and assigned to species taxa (spec.), the number of filtered features and classified total (assign.) and species (spec.) taxa as well as the median reads per classified total (assign.) and species (spec.) taxa of each sample (run I top, run II bottom, see Table 1 for reference).

ID	reads [10^6]			feature	taxa		median	
	feature	assign.	spec.	[10^3]	assign.	spec.	feature	spec.
NTC	-	-	-	-	-	-	-	-
EC	-	-	-	-	-	-	-	-
ZyDNA	0.5	0.5	0.5	4.0	152	136	9	37
ZyCell	1.0	1.0	1.0	7.9	199	184	8	50
St61	0.5	0.5	0.5	8.1	501	429	7	27
St01-1	0.6	0.6	0.5	8.9	562	484	7	38
St02-1	0.5	0.5	0.4	5.6	416	357	7	28
St01-2	0.4	0.4	0.4	5.4	401	345	7	30
St03-1	1.0	1.0	1.0	10.5	645	560	8	39
St04-1	0.7	0.7	0.7	7.5	476	411	7	26
St05-1	0.5	0.5	0.5	6.2	424	360	8	41
St02-2	0.4	0.4	0.4	4.6	403	358	7	22
NTC	<0.1	<0.1	<0.1	0.7	89	79	8	30
EC	0.2	0.2	0.2	2.0	182	165	7	26
ZyDNA	0.4	0.4	0.4	7.3	236	220	6	19
Lu05	0.1	0.1	0.1	1.6	161	142	8	24
Lu13	0.1	0.1	0.1	1.6	211	183	6	14
Lu18	0.1	0.1	0.1	1.6	229	196	6	11
EC-en	0.1	0.1	0.1	1.1	111	99	8	25
Lu05-en	<0.1	<0.1	<0.1	0.7	130	117	8	14
Lu13-en	0.1	0.1	0.1	0.8	75	62	9	33
Lu18-en	0.1	0.1	0.1	0.8	53	46	7	17
Lu05-en2	0.2	0.2	0.2	1.3	165	143	7	22
Ov85-en	0.7	0.7	0.6	5.8	320	293	7	24

The control samples retained between 0.7*10^3-2.0*10^3 features, in agreement with previous observations.

The OTU open-reference clustering successfully reduced the number of input reads to a smaller number of features.

Taxonomic Classification

The QIIME2 package includes a naive Bayesian classifier, intended for taxonomic classification of the extracted features. Unfortunately, the training of the classifier failed repeatedly, because of its high computational requirements. The features produced by the OTU picking were, therefore assigned to reference taxa using the k-mer mapping strategy evaluated in Section 2.4.2.

With a maximum of 01:57 min using 8 threads, the process was faster than the k-mr mapping of sequening reads, due to the reduced input set. Nearly the full set of remaining features was classified for all samples, showing no change between the number of clustered and assigned reads (see Table 9). Similar to the k-mer mapping of the sequencing reads, the mapping of features resulted in a high fraction of species assigned reads and species taxa (between 82% and 93%), confirming a good resolution of the methodology. At the same time, with between 46 and 560 taxa the fewer species were assigned than the for the k-mer mapping alone or k-mer mapping and reestimation. At the same time, the median number of reads per taxon was higher.

The number of taxa assigned to the mock community samples showed a reduction in complexity compared to the clinical samples, as previously described (see Section 2.4.2 and 2.4.3). The same was true for the samples enriched in batch I.

Overall, the k-mer mapping of the features showed the same results as the k-mer mapping and alignment, by retaining a high number of reads for the clinical samples and control samples of run II, but an reduced number for the enriched samples.

Mock Community

To investigate the influence of the pre-clustered input on the taxonomic classification, we compared the results of the mock community samples between the k-mer mapping, using unclustered quality reads followed by the reestimation (Section 2.4.2), to the assignment of clustered features.

The results of both approaches showed similar trends in the results, but the OTU picking was less accurate (compare Figure 13 and Figure 8). The fraction of reads aligned to taxa, other than the reference, were higher for all investigated ranks (up to 84% for species, 4% for genus and 2% for family rank) and the relative deviations

from the reference was increased for most of the taxa (up to >100 fold for species and 4 fold for genus and family rank).

A analysis of the taxonomic composition revealed a wide agreement in the classification of the most abundant taxa between both methodologies (compare Figure 14 and 9). However, the reduction of input feature set did not widely improve the classification. The same set of high abundance taxa with unspecific association were detected at species rank. Further, even though the number of low abundance species was reduced, this did not increase the assignment of reads to the correct species. Taking the genus *Lactobacillus* as an example: Besides the reference *Lactobacillus fermentum*, there were 78 more species rank taxa detected for the k-mer mapping using unclustered reads. In contrast, the same genus included only *Lactobacillus fermentum* and 26 other species rank taxa for the clustered input. However, at the same time, this genus was supported by 69'740 reads (4.77% of the total assigned reads) after k-mer mapping and reestimation of abundance, but only 38'000 reads (3.55% of the total assigned reads) when using the clustered features as input.

The described results also were reflected in the statistical measures of the community samples (see Table 10), showing an increased in evenness, as well as decreased diversity and richness, compared to the k-mer mapping of unclustered reads.

TABLE 10. Mock community properties – OTU picking: Listed are the species richness (S), the diversity as Shannon entropy (H) and Simpson index (λ) as well as Pielou's evenness (J) in all samples and the underlying community reference.

run	**ID**	S	H	λ	J
I	ZyDNA	152	2.53	0.86	0.50
I	ZyCell	199	2.67	0.86	0.50
II	ZyDNA	236	2.66	0.86	0.49
-	reference	8	2.01	0.86	0.97

Overall, the OTU picking purposely reduced the number of classified taxa in the mock community and the statistical properties were closer to the reference. This was, however, not reflected by an increase in the number of correctly assigned reads (neither relative, nor absolute).

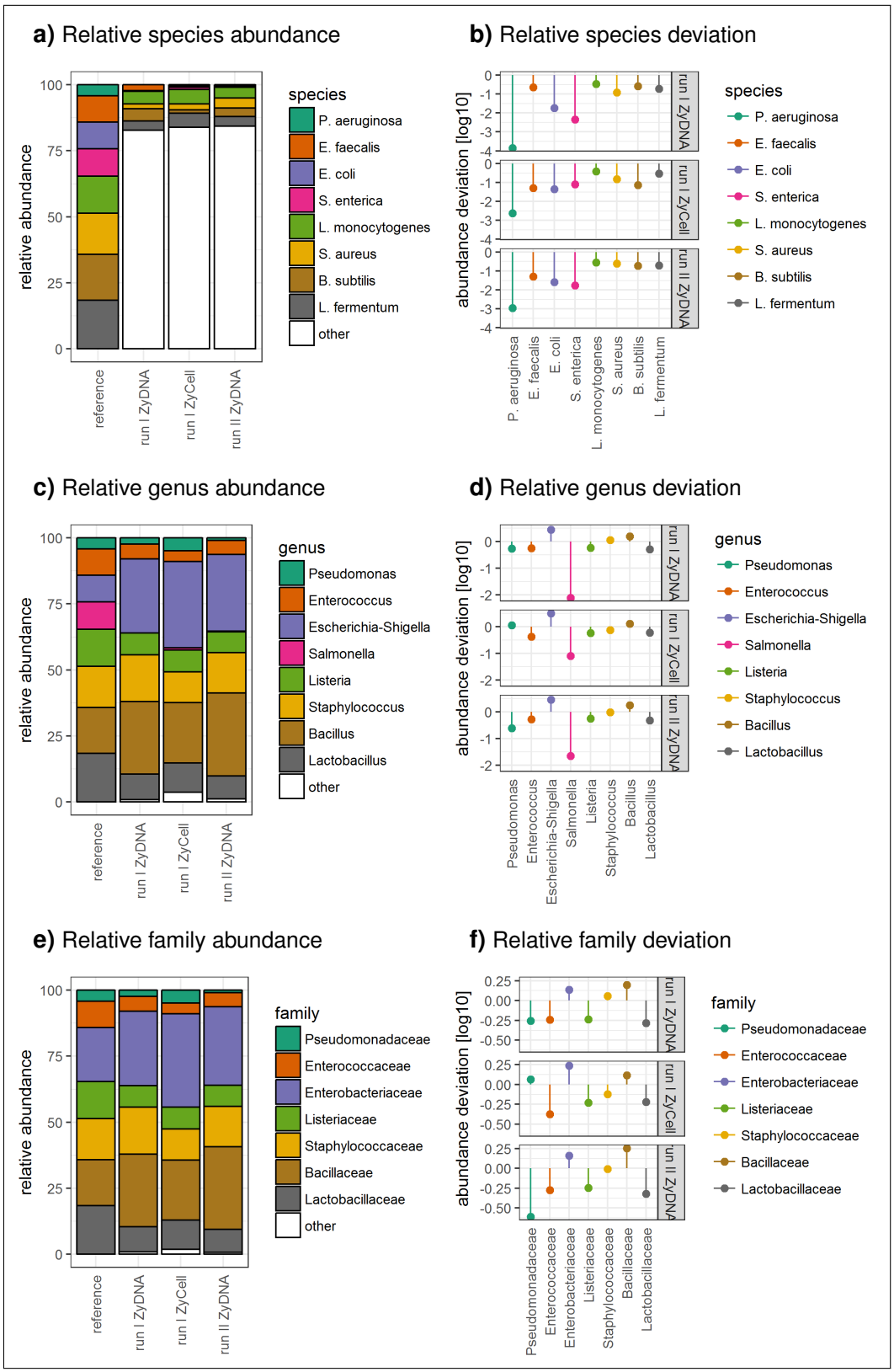

FIGURE 13. Mock community taxonomic abundance and deviation – OTU picking: Shown are the abundance (Subfigure **a)**, **c)** and **e)**) of the three mock community samples relative to the abundance in the input material (reference) as well as the relative deviation from it (Subfigure **b)**, **d)** and **f)**) for species (Subfigure **a)** and **b)**), genus (Subfigure **c)** and **d)**) and family (Subfigure **e)** and **f)**) ranks.

FIGURE 14. Mock community taxonomic composition – OTU picking: Multi-layer pie chart of the taxonomic composition for the mock community, exemplarily shown for ZyDNA of run I open-reference clustering and k-mer mapping. Slices indicate the relative abundance of the taxa at each rank (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards from higher taxa represent reads assigned to higher level taxa. Shaded slices represent multiple taxa. The taxonomic assignment of the other mock communities were not included.

2.4.5 Amplicon Sequence Variants

Recovery of amplicon sequencing variants is an alternative approach to define features in the set of quality controlled reads. An initial read cluster of similar reads had to be selected. Other reads could then be compared to identify erroneous reads originating from the same template. An error model is inferred from these associations to be able to include other reads and form other clusters with similar

properties.

The ASV approach was applied on our sequencing reads but was unable to generate the error model in both of our runs. The algorithm required an initial cluster of reads with identical sequence in order to align similar reads and extract information about the errors introduced during their creation. As discussed before (see Section 2.4.4), the properties of the Nanopore sequencing reads make it unlikely to find clusters of identical reads.

Hence, the algorithm resorted to an initiation with singleton clusters and interpreted all variability in the data set to be caused by introduced errors. The resulting error frequencies did not follow the consensus quality score (see Figure S9). This resulted in a combination of all reads into one single cluster, making it not applicable for our circumstances.

2.5 Clinical Samples

MeBaPiNa was developed for the automated analysis of patient microbiomes. To investigate this capability, we sequenced and analyzed a set clinical samples alongside the mock communities and control samples. As described before (see Section 2.4.3), the full-length alignment achieved a higher resolution of taxa at the species rank. We, therefore, used the results of this methodology for further analysis.

2.5.1 Control Samples

Before the investigation of the clinical samples, the control samples were analyzed to investigate the potential of a kit contamination in the sequencing runs.

As described in Section 2.4.3, the control samples of run I had only a small number of assigned reads. All reads of the extraction control were excluded and the remaining ten reads of the no-template-control sample were assigned to three taxa of genus *Staphylococcus* and *Burkholderia-Caballeronia-Paraburkholderia*. The observed characteristics suggested that the reads did not represent a contamination, but originate from systematic errors or wrong barcode assignment.

The taxonomic classification of the NTC sample of run II showed a number of classified taxa of different genera (see Figure S10). A high fraction of *Burkholderia*

species from the genus *Burkholderia-Caballeronia-Paraburkholderia* were detected in the sample, comprising 74% of the assigned reads. It remained unclear whether a mixture of species or a bad alignment was responsible for the species diversity in this genus. The remaining reads were associated with *Moraxella osloensis* (21%) and *Granulicatella adiacens* (4%), as well as other low abundant species.

The taxonomic classification of the extraction control was comprised of a large fraction of species from the *Enterobacter* genus with 43% of the assigned reads (see Figure S11). The largest single classified species was *Moraxella osloensis* (19%), analog to the no-template control. Most of the remaining reads were classified as taxa contained in the *Staphylococcus* (15%), *Paracoccus* (8%) and *Pseudomonas* (6%) genera.

The results suggested the amplification of kit contaminations in run II with a small number of genera in both the extraction reagents and the kits used for library preparation. This contamination has to be accounted for, when further analyzing the samples.

2.5.2 Gut Microbiome

Run I was comprised of clinical stool samples, used to analyze the gut microbiome of the patients.

The samples showed overall comparable properties (see Table 11). The observed differences in richness are likely caused by the difference in the number of quality controlled input reads. All samples had higher diversity and lower evenness than the mock community (compare Table 8), reflecting the differences between an artificial community with few input species and a patient sample with higher complexity.

The taxonomic distribution of the stool samples showed a large fraction of taxa in the order of *Lachnospirales* or *Oscillospirales* (see exemplary Figure 15). Besides the *Firmicutes*, the samples included mainly *Bacteroidota* and *Proteobacteria*. The overall high diversity was represented through a diverse number of species and genera. Species with high abundance were detected alongside a large number of species with low abundance and the assigned taxa showed overlaps between the samples, as well as sample specific assignments (not quantified). Many of the

TABLE 11. Gut microbiome properties: Listed are the species richness (S), the diversity as Shannon entropy (H) and Simpson index (λ) as well as Pielou's evenness (J) in all samples and the underlying community reference.

ID	S	H	λ	J
St61	447	4.02	0.96	0.66
St01-1	526	4.49	0.98	0.72
St02-1	374	3.82	0.96	0.64
St01-2	298	3.46	0.93	0.61
St03-1	575	4.47	0.98	0.70
St04-1	445	3.76	0.94	0.62
St05-1	384	3.94	0.96	0.66
St02-2	336	3.87	0.96	0.67

classified taxa had unspecific names assigned, some of those still reflected the sample origin, e.g. "human gut" or "gut metagenome".

The analysis pipeline resulted in a plausible representation of the microbial communities in the clinical samples. The higher diversity is in contrast to the mock community with lower complexity.

Time Course

The pipeline was developed as part of the PROMISE trial in order to find predictive markers for the treatment outcome. As such, the intent is to apply further analysis on the classified taxa, outside of the scope of my work, in order to analyze the temporal dynamics of the microbiome. Four samples of two patients taken before and after the treatment initiation were included in the samples set to evaluate the feasibility of this future work.

Patient 1-001 showed a strong reduction of richness (from 526 to 298) and diversity (Shannon entropy from 4.49 to 3.46) after initiation of the treatment (see Table 11 and compare Figure 15 and Figure 16). A drastic decrease of *Oscillospirales* order taxa from 25% to 3% of the assigned reads was observed. Further, many of the lower rank taxa were excluded, leaving a smaller number of taxa with higher fractions.

In contrast, the composition of patient 1-002 did change only slightly (see Table 11). The richness decreased slightly from 374 to 336 taxa, but the Shannon entropy shifted only slightly from 3.82 to 3.87 and the evenness from 0.64 to 0.67.

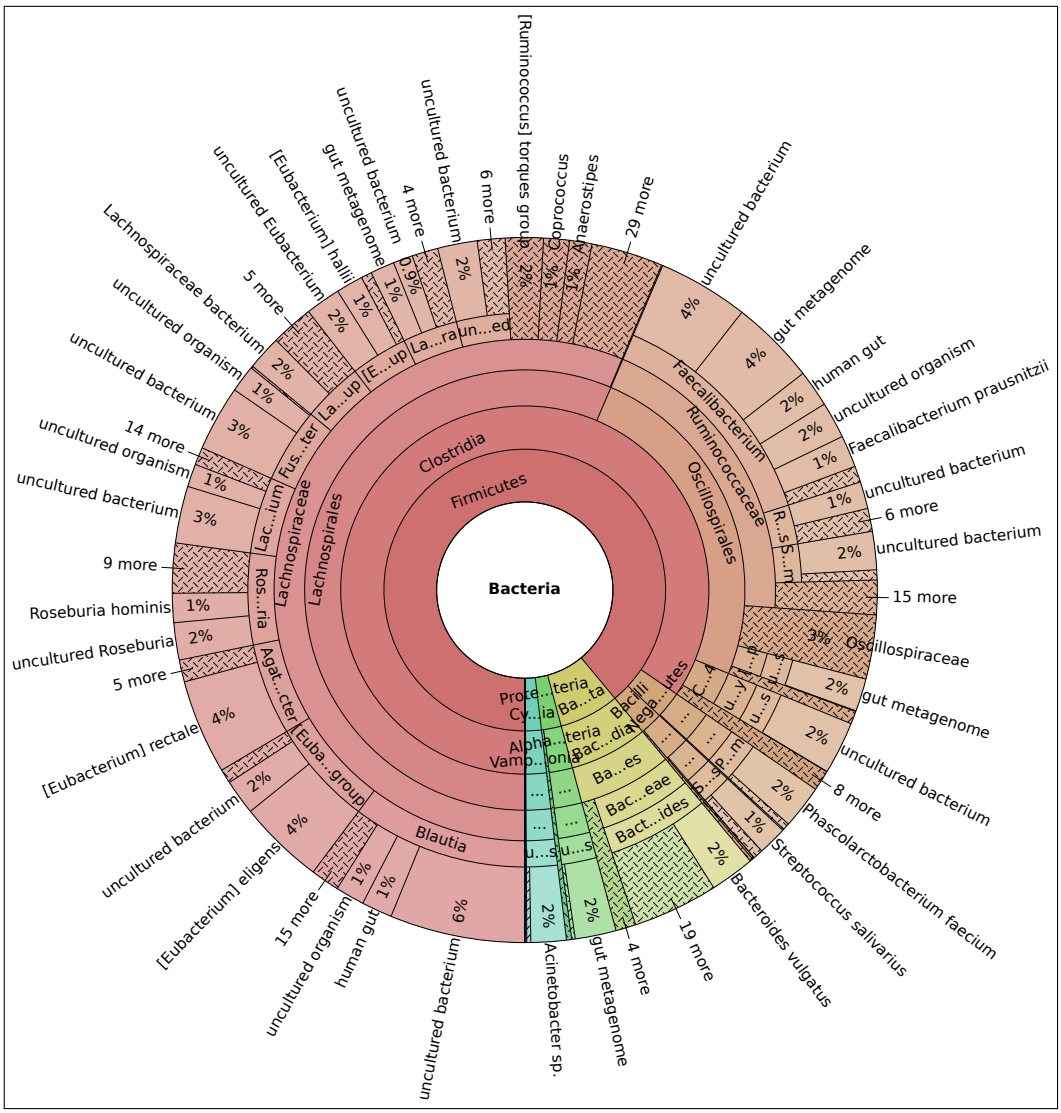

FIGURE 15. Gut microbiome taxonomic composition – patient 1-001 T1T2: Multi-layer pie chart of the taxonomic composition for the gut microbiome, shown for St01-1 using full-length alignment. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

The taxonomic compositions of this patient did not drastically change over time (compare Figure 17 and Figure 18). In contrast to the large differences seen between the samples of patient 1-001, the distribution in higher rank taxa remained comparable between the timepoints. Similarly, both timpoints showed a large variety in the abundance of taxa, including low abundance and high abundance species. However, besides the overall comparable diversity, the composition of genera and

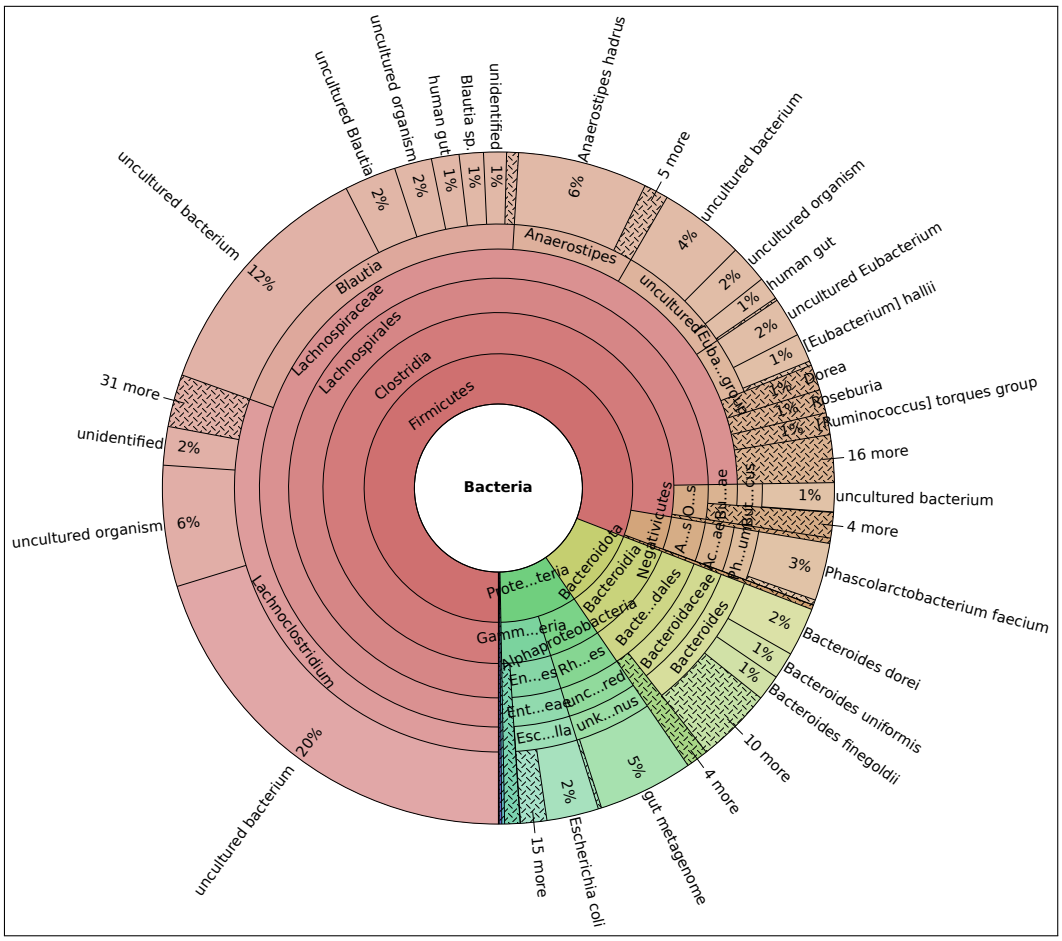

FIGURE 16. Gut microbiome taxonomic composition – patient 1-001 T2ER: Multi-layer pie chart of the taxonomic composition for the gut microbiome shown for St01-2. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

species changed, e.g. the 100 fold increase of *Bacteroides fragilis* from 0.09% of the reads before treatment to 9% at the second timepoint or *Roseburia intestinalis* from 0.7% to 11%, which could resemble natural fluctuation of the gut flora.

The examined samples showed drastic differences in the dynamics of their microbial composition between both time points, represented by differences in the statistic parameters and the taxonomic composition. These results were of potential value in the search for predictive markers of the treatment outcome.

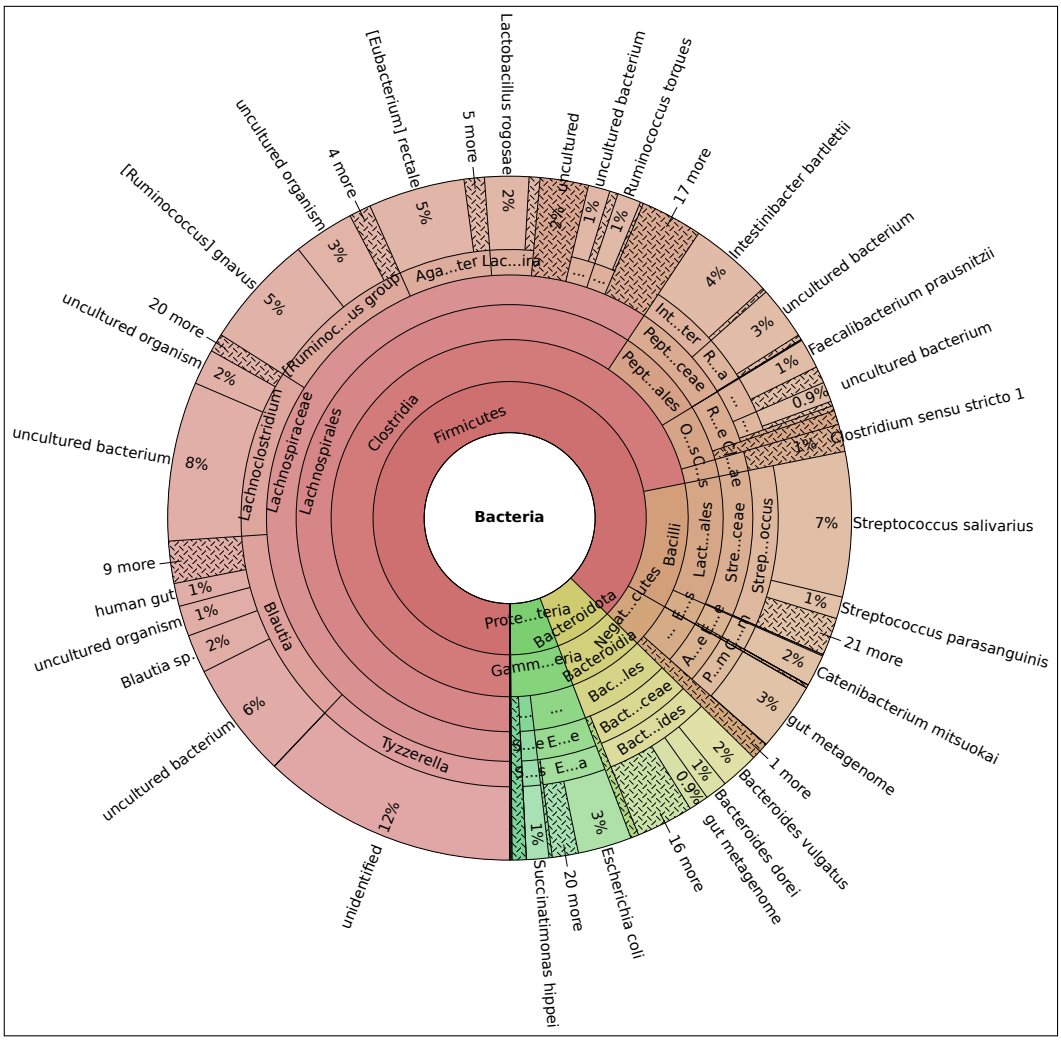

FIGURE 17. Gut microbiome taxonomic composition – patient 1-002 T1T2: Multi-layer pie chart of the taxonomic composition for the gut microbiome, shown for St02-1 using full-length alignment. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

2.5.3 Tissue Microbiomes

A further target of investigation was the tissue microbiome. Since the analysis of low biomass is complicated by the small amount of input DNA and the higher amplification of kit contaminants, we investigated the capability of the pipeline to investigate the microbial composition of four different tissue samples and the effects of enrichment.

FIGURE 18. Gut microbiome taxonomic composition – patient 1-002 T2ER: Multi-layer pie chart of the taxonomic composition for the gut microbiome, shown for St02-2 using full-length alignment. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

The tissue samples show lower richness, diversity and evenness then the stool samples (compare Table 11 and 12), which were, however, sequences in two individual runs making a direct comparison unreliable. However, the unenriched lung samples were comparable in terms of diversity and envenness to the mock community of the same run, suggesting a lower complexity in the samples. The same was true for the ovary sample.

TABLE 12. Tissue microbiome properties: Listed are the species richness (S), the diversity as Shannon entropy (H) and Simpson index (λ) as well as Pielou's evenness (J) in all samples and the underlying community reference.

ID	S	H	λ	J
Lu05	184	2.82	0.89	0.54
Lu13	103	2.31	0.82	0.50
Lu18	150	2.60	0.85	0.52
Lu05-en	79	2.46	0.84	0.56
Lu13-en	61	2.18	0.82	0.53
Lu18-en	39	1.31	0.59	0.36
Lu05-en2	22	1.20	0.60	0.39
Ov85-en	353	3.03	0.89	0.52

Exemplary analysis of the taxonomic composition of the not enriched sample of Lu05 (see Figure 19) showed a composition similar in complexity to the mock community. Few highly abundant species of the *Escherichia-Shigella* genus were covered by 41% of the reads. The rest of the reads was separated between *Alphaproteobacteria* as well as *Bacilli*. A more diverse mixture of taxa with low and high abundance was detected for these subgroups.

Besides the taxa mentioned above, 18% of the reads were assigned to the *Moraxellaceae* family. As described in Section 2.5.1, the same taxa was detected in both of the control samples and has to be excluded in the analysis of this run II.

As described before, the ovary sample was comparable in diversity to the unenriched lung samples. This was further supported by the taxonomic composition shown in figure 20. A large number of taxa were detected in the sample and a large similarity to the lung sample was observed. The largest fraction of reads was also assigned to *Escherichia-Shigella* genus, but a large variety of other *Cyanobacteria* and species of genus *Enterobacter* were detected in the ovary sample, but not the lung sample and the genus and species level assignment differed in many taxa. Further, the relative fraction of *Moraxellaceae* was much lower in this sample. Likely due to the higher overall coverage.

An overall diverse set of taxa was detected in both types of tissue. A similar set of taxa was detected in both tissue types. However, the composition of the other patients was different (not shown).

FIGURE 19. Tissue microbiome taxonomic composition – patient Lu05 not enriched: Multi-layer pie chart of the taxonomic composition for the lung microbiome, exemplarily shown for patient Lu05 using full-length alignment. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

Enrichment

As previously noted during the taxonomic classification (Section 2.4), the enriched samples showed lower richness than the same samples before enrichment, in all lung tissue samples. This was reflected by a reduction in diversity and evenness measurements in some, but not all of the samples (see Table 12).

FIGURE 20. Tissue microbiome taxonomic composition – patient OV85 enriched: Multi-layer pie chart of the taxonomic composition for the ovary microbiome, shown for patient Ov85 using full-length alignment. Slices indicating relative quantity of reads assigned to the taxa ranks (from inside out: phylum, class, order, family, genus, species). Slices reaching outwards represent reads assigned to higher level taxa. Shaded slices represent multiple taxa.

Further analysis of the example of patient Lu5 showed drastic changes between the sample before (Figure 19) and after enrichment (Figure S12 and Figure S13). Both enriched samples showed a large fraction of *Moraxella osloensis*, likely originated from the kit contamination seen in the control samples. Further, *Burkholderia* were detected in the enrichment sample Lu05-en, but not the not enriched sample, which likely also originated from the contamination. Comparing the taxonomic composition

of the remaining taxa, a clear decrease in the number of taxa and species with low abundance upon enrichment was visible. This trend was even stronger in the sample of enrichment batch II than in the sample of enrichment batch I. Many taxa (e.g. *Enterobacterales*), strongly represented in the original sample, removed upon enrichment. Other taxa showed an strong enrichment or were not detected before the enrichment (e.g. *Stenotrophomonas*). No clear pattern could be observed. Similar trends were visible for the other lung samples (data not shown).

Overall, the enrichment skewed the classification of the analyzed microbiomes and reduced richness and diversity instead of increasing it.

3 Discussion

The aim of this work was to develop an automated pipeline for the metabarcoding analysis of clinical samples. As part of the PROMISE trial this provides a fast and reproducible tool to analyze a large cohort of patient samples of multiple sources, to extract biomarkers capable of predicting treatment outcomes.

3.1 Sequencing Results

The chosen clinical samples included a set of different microbiomes to represent the complexity of biomaterials collected during the PROMISE trial (PROMISE trial consortium 2018). The gut microbiome was represented by a collection of six different patients, including multiple time points for two of them, to investigate the dynamic between different time points of the treatment course. The tissue microbiome was represented by another four samples of different sources. During the analysis of the PROMISE trial itself, the samples of both sources can be compared for each patient to get an even higher level understanding of the alterations and their connection (Enaud et al. 2020). Run I and run II, performed by Dr. med. Silke GRAULING-HALAMA, both included mock community samples as well. These allow the quantitative comparison of results to assess the performance of the different methodologies and facilitate the comparison of the runs.

Both sequencing runs were completed successfully and resulted in a high number of reads, most of which were successfully basecalled. This resulted in a throughput of 18.2 Gb in 48 h for run I, which is within the expected range specified by the manufacturer (Oxford Nanopore Technologies, Limited. 2019b), and 8.3 Gb for run II, which is lower due to the shorter run duration. The quality score of run II was lower than of run I, which could lead to higher counts of wrongly assigned reads in this run, but can be addressed during filtering.

A difference in the number of reads between the two runs was detected, which could not be explained by the difference in run duration alone. Further, a strong difference in the read length distribution of both reads was observed. Whilst run I showed

a narrow peak of reads with a length close to the expected amplicon size, run II had a large subgroup of shorter reads accounting for about half of its throughput. Sequencing throughput is dependent on the input fragment length, because at the same input mass shorter reads have a higher molarity of read ends ligated to sequencing adapters, which results in preferential sequencing of shorter reads (Byrne et al. 2019; Kono and Arakawa 2019). Hence, the difference in read length was the main reason for the difference in the number of reads between both runs, while showing a comparable number of produced bases relative to the run duration and under consideration of the decreased throughput over time.

During demultiplexing, reads were assigned to all barcodes and all, but control samples of run I were covered by at least 0.1*$*10^6$ reads. The absence of coverage for the control samples of run I , but not run II, suggests that the dominance of kit contaminations is strong in run II and negelectable in run I, which will be discussed further in Section 3.4. Since the number of reads per non-control sample does not show a link to the input mass, the source of the coverage differences between the other samples remains unclear. Likely it is caused by the accumulation of systematic and stochastic errors (e.g. degradation of the fragments or pipetting errors) leading to the variable representation of the fragments in the library. A large fraction of reads was also excluded due to the missing detection of a barcode sequence above the preset threshold. There is, however, no reason to believe that these reads belong to one specific barcode, as their quality distribution indicates an overall high error rate, which implies a higher probability of errors in the barcode sequence, independent of their origin. The barcode detection threshold could be reduced to include these in the pool of analyzed reads, but this has the potential to lead to false classifications while increasing the processing requirements. The low fraction of reads assigned to the control samples of run I also indicate a correct barcode detection, as no or minimal samples-bleeding was observed. However, sample-bleeding depends on the sequence similarity between pairs of barcodes (Wright and Vetsigian 2016) and cannot be completely ruled out for other barcode pairs.

Closer inspection of the read length distribution of run II revealed several further small clusters of reads, in between both main clusters and one above the expected amplicon size. These patterns are unlikely to occur from degradation in the library or incomplete sequencing, as this would not lead to the formation of defined clusters, but to the tail of decreased read length seen in run I. Further, this reasoning could

not explain the cluster above the expected read length. Since this phenomenon is limited to the tissue samples of run II, but not the controls or mock community, the sample origin and the increased number of amplification cycle in run II are likely causal factors. An brief analysis of a small subset of the reads from the largest secondary cluster (BLASTn search with default parameters) revealed an overlap to the 18S rRNA of several species (data not shown). In this context, a primer mismatch leading to the unwanted amplification of a 18S rRNA fragment is most likely, since both genes are closely related and can even be amplified together deliberately (Klindworth et al. 2013; Wang et al. 2014). The increased number of 45 PCR cycles in run II and the preferential selection of short reads in Nanopore sequencing strongly amplified this signal leading to the observed read length distribution. The exact origin of the amplicons remains unclear. However the large fraction of host cells in the tissue samples suggest a human origin.

The clusters of read lengths in run II are far enough apart to enable a size selection via a bead clean-up during library preparation (DeAngelis et al. 1995). This is preferred over bioinformatical filtering, as it does not reduce the sequencing throughput. Nevertheless, the filtering using length and quality filters was successful in removing the subset of short reads, low quality artifacts and any chimeric reads outside the filter range, while leaving a subset of reads in the range of the expected amplicon size. Despite the filtering, the remaining main read length peak still sowed a width of about 100-200 bp. This spread in read length can be explained by the variability in the lengths of 16S rRNA gene between species. Analysis of our reference database showed a mean length of 1440 bp with standard deviation of 132 bp which agrees with the distribution we observe for the trimmed reads. The additional length of the untrimmed reads originated from the adapter and barcode sequences at both read ends.

The pipeline involves predefined thresholds for length and quality. The length threshold was chosen to include the expected amplicon size, while accounting for partially sequenced or fragmented reads. The quality threshold was based on the default suggestion of ONT (Oxford Nanopore Technologies, Limited. 2018a), which is widely used, but has no reasoning. The read filtering could also be skipped, as the subsequent analysis steps include further filtering, e.g. by removing all reads without alignment or taxonomic classification. However, the filtering remains a reliable way to remove low quality reads which could lead to misclassifications and an artificial

increase of diversity and would include further computational requirements. A future extension of the pipeline could use data driven approaches to determine thresholds based on the quality and reads length distribution.

With a median read accuracy of 93%, the error of the filtered is in a range with previously reported studies (Jain et al. 2017), but still almost two orders of magnitude higher than the error rates achieved by second-generation sequencing (Rang et al. 2018). Nevertheless, the superior read length covers a larger set of variable regions, counteracting this effect.

3.2 Analysis Methodologies

The different methodological approaches were compared to identify the approach best suited for the analysis of the clinical samples. A mock community of known composition was used to assess the performance of the individual methodologies. This allowed the quantification of the method accuracy in terms of deviation from the reference as well as the comparison to previously reported results (i.e. Benítez-Páez and Sanz 2017; Cuscó et al. 2019; Edwards et al. 2016).

The feature extraction using OTU picking was successful under an identity threshold of 85%, as reported by Curren et al. 2019. However, even at a 97% cutoff, the species diversity can be underrepresented (Callahan et al. 2016; Edgar 2017) and the applicability of feature extraction methods for Nanopore sequencing remains questionable, especially for the superseded OTU picking. The applicability, was further questioned by the unsuccessful attempted of ASV recovery. The generation of the required error model failed, because no read clusters could be identified, suggesting that the error rates are to high for feature extraction and explaining the absence of reports of ASV recovery for Nanopore sequencing. An unofficial statement form the main developer of DADA2 confirmed, that the fraction of error-free reads has to be at least about 10% in order to successfully recover ASVs (Callahan 2019).

Instead of feature extraction, the sequencing reads can be used as individual features themselves, as demonstrated using k-mer mapping and full-length alignment. The direct comparison confirmed the reduction in complexity through OTU picking, which did not lead to an increase in the number of correctly assigned

reads and showed the highest deviation from the reference. This might be caused by the ambiguities that has to be solved during the clustering process, where reads with equal distance to existing clusters have to be assigned to one or the other, despite the incomplete information (Rognes et al. 2016).

A comparison of k-mer mapping and full-length alignment showed a good overlap of both methods with the reference at higher level taxa, but a higher fraction of reads assigned to the correct reference species taxa was observed for the full-length alignment. Both methodologies showed a large number of low abundance taxa besides the main groups, but the number of classified taxa was lower for the alignment approach. It has to be noted that the confidence threshold for k-mer mapping was set to zero to allow a high number of species assignments, resulting in the classification of a large number of species taxa, many of which got removed during reestimation of abundance. This could explain the formation of the additional low abundance assignments and can be prevented by an increase of the confidence threshold. Since the alignment approach used more stringent filtering, a difference in the number of assigned taxa was expected. This was also reflected by a reduction in the number of overall assigned reads for the full-length alignment. Similar to the quality control steps, the selection of the appropriate filtering parameter is not trivial. A trade-off between a higher number of potentially false positive and false negative classifications has to be made. Further empirical tests could help to optimize the parameter set for our application.

Overall, the results showed a large agreement with the reference community and a high reproducibility between the runs. Comparison of the results of our classification to other published results, were difficult because of the differences in the experimental setup. Edwards et al. (2016) observed much stronger deviations, even at higher taxonomic ranks, likely due to the earlier generation of software and hardware used. The species rank classification reported by Cuscó et al. (2019) are superior to the results reported here. However, the 16S rRNA reference database used in this publication was limited to the expected reference taxa, making a direct comparison impossible. Benítez-Páez et al. (2016) observed species deviations in a similar range to the observation we made, but used a different mock community.

3.3 Classification Biases

The largest deviations from the reference were seen to be caused by reads being assigned to taxa with missing species nomenclature (e.g. "metagenome" or "uncultured bacterium"), particularly in the case of k-mer mapping. Since the taxonomic assignments of the SILVA reference database did not include species rank classifications (Yilmaz et al. 2014), the species assignments of the RefSeq database (O'Leary et al. 2016) were used. Unfortunately, this includes a large number of ambiguous assignments as seen in the taxonomic classification results reported here. The associated sequences could be filtered out during generation of the reference database. This would, however, force any reads represented by this reference to be assigned to an alternative sequence (Cuscó et al. 2019), potentially with a lower confidence, leading to an overall decrease in diversity and assignment confidence. The alignment to closely related species could also be a possible explanation for the increase in error rates between the read Phred score and alignment identity.

Generally, a reference data base can be freely chosen and several alternatives to the SILVA database are available, including Greengenes (McDonald et al. 2011), RDP (Cole et al. 2014), NCBI RefSeq (O'Leary et al. 2016; Tatusova et al. 2014) and others. Whilst all listed references are valid choices, they differ strongly and no consensus on choosing one over another could be found (Balvočiūtė and Huson 2017; Schloss 2009; Werner et al. 2011). Whilst the RefSeq database contains the largest set of reference sequences of the listed resources, it contains many duplicates and is not as well curated as the other databases. On the other hand, the latest update in Greengenes was in 2013, leaving out important changes and additions to the taxonomic tree (Balvočiūtė and Huson 2017). SILVA was selected because of its recent update (2019) and manual curation process, but the inclusion of the RDP and RefSeq database as alternative references into the database would be an desirable addition.

Overall, the selection of a reference database remains a trade-off between completeness and curation. In this regard, the feature clustering approaches are advantageous as they can identify clusters without matching sequence in the reference database if applied as de-novo clustering or ASV recovery (Hugerth and Andersson 2017). This advantage could be used to reanalyze reads discarded

by full-length alignment or k-mer mapping to identify clusters of sequences not represented in the reference database.

Another source of missing or reduced species counts emerges already during sampling and library preparation. An analysis with SILVAs TestPrime (Yilmaz et al. 2014) revealed that only about half of the bacterial sequences in the database are covered by the primer pair utilized in the 16S sequencing kit of ONT, with one allowed mismatch. This is mainly caused by the low coverage of the reverse primer (Klindworth et al. 2013). Further, DNA extraction is dependent on the efficiency with which the bacterial cells are lysed. This on the other hand strongly depends on the properties of their cell walls, as determined by the Gram staining (Gorzelak et al. 2015; McOrist et al. 2002; Walker et al. 2015). The tested mock community included three easy to lyse Gram-negative bacteria and five Gram-positive bacteria which are harder to lyse. However, the results observed by the analysis of the mock community samples in our experiments revealed no correlation between the Gram stain and the deviation of the ZyCell sample from the reference, which indicates an efficient lysis through bead beating. Furthermore, the results of the DNA based mock community agreed with the results of the cell sample, further confirming the efficient lysis. The GC content of the amplified region can also influence the classification results, as GC-rich sequences tend to be inefficiently amplified in PCR reactions (Browne et al. 2020). The results of our analysis did however not show a dependence of the deviation on the GC content reported by the manufacturer, either.

It remains unclear whether the low complexity and relatively large abundance of species in the mock community is a good representation of real microbiota. The results from the mock community suggest that stronger filtering should be applied to reduce the large number of taxa with low abundance. However, in clinical samples this would likely lead to a strong reduction of diversity and loss of detection of low abundance species.

3.4 Clinical Samples and Kit Contamination

Our results indicated the presence of a contamination in the reagents used for the DNA extraction as well as the library preparation. It is widely known that DNA contamination are present in laboratory reagents (Salter et al. 2014; Weiss

et al. 2014) and that their influence can lead to biased results. Salter et al. (2014) also describes that these contaminations are especially problematic for low biomass samples because of the required increase in amplification and the associated increase of the contamination. As we can see in comparison of run I and run II, this was also the case for our setup. Run I, which had a high amount of microbial input and only 25 amplification cycles, does not show a high number of classified taxa in the control samples, which were even fully excluded during the processing with some of the methodologies. In contrast, the control samples of run II showed a large number of classified taxa in both control samples. Both showed independent populations of different *Gamma-proteobacteria* and *Firmicutes*, reflecting the contaminations described by Salter et al. (2014). It is important to consider these contaminations when assessing the diversity and composition of the clinical samples. A downstream analysis of the classified taxa should include a step to partially or fully exclude reads assigned to the corresponding taxa from the sample, to prevent a false positive classification.

The microbiome has been shown to have important implications for many disease conditions (Falony et al. 2019; Pflughoeft and Versalovic 2012; Wang et al. 2017), including the treatment of cancer (Gopalakrishnan et al. 2017; Matson et al. 2018; Routy et al. 2017). The PROMISE trial aims at the detection of biomarkers to predict the treatment outcome in patients with NSCLC. The set of analyzed biomarkers includes samples of gut and tissue microbiota and their tracking over time.

The established pipeline was used to showcase the classification of samples from these sources as well as the ability to investigate changes between different time points. The results of the gut microbiome samples were in alignment with the higher taxa reported in the literature (Rinninella et al. 2019) but also showed patient specific compositions and especially dynamics over time. A general underrepresentation of *Bacteroidota* was observed, which can be linked to the biased primer coverage (described in Section 3.3) of only about 10% for species of this phylum. While the microbial composition of patient 1-002 showed mostly changes at lower level taxa, the diversity of patient 1-001 showed a strong decrease between the time points. The sample set is to small to allow further analyses, but the results indicate that the detection of patient specific changes over time is possible.

The composition of the lung tissue samples showed a lower complexity than the stool samples and was in agreement with the reported reference (Moffatt and Cookson

2017). *Proteobacteria* and *Firmicutes*, detected in the ovary sample, showed a ratio similar to the reference literature (Zhou et al. 2019), but the detected *Cyanobacteria* did not fit this description and might indicate a patient specific alteration or an undiscovered disease condition. Over all tissue samples from run II, a presence of the taxa from the control samples was detected. This illustrates the need for an additional processing to exclude these false positive classifications, since a reduction of the amplification cycles is not possible for low biomass samples.

The effect of the contamination was strongest for the enriched samples showing large fractions of reads assigned to the taxa detected in the control samples. Together with the large decrease in the number of classified taxa in the enrichment samples, this suggests a depletion of the sample DNA during the enrichment and subsequent amplification of the contamination during the library preparation. The results show that 16S rRNA does not benefit from sample enrichment, since the sequence specificity of the primers alone is sufficient to largely deplete host contaminations in the form of human DNA. The same was demonstrated by spike-in contaminations of human DNA (Kai et al. 2019).

3.5 Pipeline Requirements

In order to automate the analysis and integrate the set of available analysis methodologies I developed MeBaPiNa. Three of the four methodologies were successfully implemented. However, ASV recovery could not be completed because of the increased error rates. The pipeline includes a configuration file to select the desired methodology and specify important parameter.

The snakemake workflow management system was selected as foundation of the pipeline as it ensures the reproducibility by including the required package dependencies as well as the code for execution of the processing steps. As dependencies and execution order of the steps is determined automatically, this also allows to rerun certain steps of the pipeline without the need to manually select and execute the required tools. Basecalling and quality control as well as automated production of visualization and statistics to provide a full support from raw sequencing input to taxonomic classification. Finally, the pipeline allows for parallel

processing of all multiplexed samples of one run, although an extension to multiple runs would in principle be possible.

Another important aspect was the utilization of free and available software. This is important so that changes in the raw sequencing input do not lead to incompatibilities of incorrect results. For large parts of the pipeline this could be fulfilled but the full-length alignment uses a custom script to collapse and filter the reads during the taxonomic classification. Further, the utilization of available tools for visualizations has the disadvantage of limited configurability. Most notably was this for the Figures 4, 5, 6 and 7, produced by NanoPlot (De Coster et al. 2018). The sample labels on the x-axis had several weaknesses and the coloring was unnecessary. Further, most of the visualization tools are limited to a report in html format. This way it is difficult to extract the required figures.

3.6 Conclusion and Outlook

We produced a pipeline to automate the analysis of metabarcoding datasets produced by Nanopore sequencing and to illustrate its application in the analysis of clinical datasets. The pipeline was successfully implemented and the analysis of the mock community showed an agreement with the reference and other publications. The exemplary analysis of clinical datasets showed a taxonomic composition representative for the sample source and revealed patient time point specific alterations.

Most of the results discussed are influenced by biases like contamination or strict filtering. However, when considering the larger aim of the project, to compare different time points and donor cohorts, the absolute abundance of a taxon is not in the focus of interest. Instead, relative changes of taxa have to be analyzed. This is comparable to the analysis of transcriptomic datasets (Hugerth and Andersson 2017). Hence, the taxonomic assignments can be analyzed with the packages developed for this task, like DESeq2 (Love et al. 2014) or edgeR (Robinson et al. 2009). This way, the batch effects can be corrected and the library size can be normalized.

4 Materials and Methods

4.1 Sequence Generation

The generation of the sequencing data was performed by Dr. med. Silke GRAULING-HALAMA, in parallel to my work. The following sections give an overview of her work, with a focus on the properties relevant for the analysis of the data.

4.1.1 DNA Extraction

As the patients collected the stool samples at home the stool samples were stabilized in Stratec Stool Stabilizer (Stratec Molecular, Berlin, Germany) on the day of the patient´s next visit in the clinic respectively the day before and stored at -80°C after arrival at the laboratory.

The DNA of patient Lu5 and Ov85 in extraction batch I (sample Lu05-en2 and Ov85-en, see Table 1) were extracted using the QIAamp® PowerFecal® DNA Kit (QIAGEN GmbH - Germany, Hilden), following the supplied protocol. [250 mg] tissue were used as input. Bead beating was performed for 10 min. The supernatant was collected, DNA was extracted and washed following the manufactures instructions.

The ZymoBIOMICS™ DNA Miniprep Kit (D4300T, Zymo Research Europe GmbH, Freiburg) was used to extract the samples of batch II-IV (see Table 1). 1 ml of stabilized stool sample or 10-30 mg of tissue sample were used. The tissue was incubated in Proteinase K overnight at 55°C as recommended by the manufacturer following the protocol "Tissue and Insect Samples", Variante A with Proteinase K. Bead beating was performed according to protocol for 5 min. The supernatant was collected, DNA was extracted and washed following the *Feces and All Non-Soil Samples* instructions of the manufacturer. An extraction control of 200 μl Microbial DNA free water (QIAGEN GmbH - Germany, Hilden) was created, following the same steps, for both runs.

ZymoBIOMICS™ Microbial Community Standard (D6300, Zymo Research Europe GmbH, Freiburg) is a cell based mock community of known quantity listed in

Table 13. DNA from this community was extracted alongside the sample of extraction batch III, as described above.

TABLE 13. Composition of the Mock Community Standards: The relative composition of microorganism in the ZymoBIOMICS™ Microbial Community Standard and the ZymoBIOMICS™ Microbial Community DNA Standard, measured as the percentage of rrn operons in the total population as well as their Gram staining (Gram), genome size (gen), GC content (GC) and the number of rrn operons (rrn).

species	fraction [%]	Gram	gen [Mb]	GC [%]	rrn
Pseudomonas aeruginosa	4.2	neg	6.79	66.2	4
Enterococcus faecalis	9.9	pos	2.85	37.5	4
Escherichia coli	10.1	neg	4.88	46.7	7
Salmonella enterica	10.4	neg	4.76	52.2	7
Listeria monocytogenes	14.1	pos	2.99	38.0	6
Staphylococcus aureus	15.5	pos	2.73	32.9	6
Bacillus subtilis	17.4	pos	4.05	43.9	10
Lactobacillus fermentum	18.4	pos	1.91	52.4	5

4.1.2 Enrichment of Microbial DNA

Enrichment of microbial DNA was performed for the samples of tissue source as listed in Table 1.

The enrichment was performed using the NEBNext® Microbiome DNA Enrichment Kit (E2612, New England Biolabs GmbH, Frankfurt am Main), according to the manufactures instructions. Briefly, human DNA which is methylated is bound to bead-coupled MBD-Fc protein and microbial DNA is collected with the supernatant.

4.1.3 Library Preparation

DNA purity was checked by measurement of the A260/280 and A260/230 ratio on a NanoDropTM 2000 Spectrophotometer. In cases of insufficient DNA purity an additional clean-up with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, USA) was performed.

ZymoBIOMICS™ Microbial Community DNA Standard (D6305, Zymo Research Europe GmbH, Freiburg) is a mixture of purified microbial DNA of known quantities

(see Table 13). A sample of this mock community was used alongside the clinical samples during library preparation (see Table 1) of both runs.

The native or enriched genomic DNA was used to prepare the barcoded 16S amplicon library. The library preparation was performed using the SQK-RAB204 kit (Oxford Nanopore Technologies, Limited., Oxford, UK) and followed the associated protocol (Oxford Nanopore Technologies, Limited. 2019a) using LongAmp Taq 2x master mix (New England Biolabs GmbH, Frankfurt am Main). The workflow is illustrated in Figure 21.

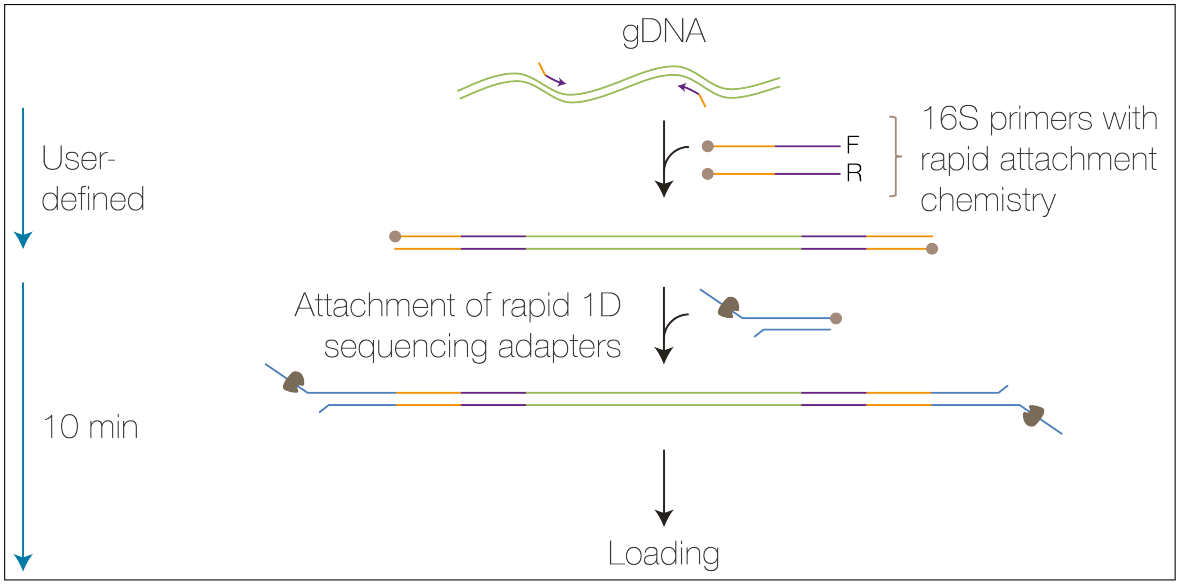

FIGURE 21. Library preparation workflow: Representation of the steps performed during library. Figure taken from ONT 16S Barcoding Kit protocol (Oxford Nanopore Technologies, Limited. 2019a).

Briefly, the gDNA of each sample was adjusted to a concentration of 1 $ng/\mu l$ and a volume of 10 μl. The barcoded forward and reverse PCR-primer (listed in Table 14) were used to amplify and barcode the 16S rRNA gene sequence of the individual samples. The PCR set up followed the specified conditions, but included 25 cycles for the samples of run I and [45] cycles for run II. For both runs, a no-template control was with using Microbial DNA free water (QIAGEN GmbH - Germany, Hilden). A subsequent bead cleanup and size selection was performed according to protocol.

The recommendation for multiplexing was an equimolar combination of all samples with a final molarity of 50-100 $fmol$ in a volume of 10 μl. The molarity of the

TABLE 14. 16S rRNA primers: The barcoded 16S rRNA gene primers consisting of an adapter, barcode and 16S specific primer sequence. Show are the forward (fwd) and reverse (rev) in 5' to 3' direction. The barcode sequence is exemplarily shown for barcode01. The position denoted M is either A or C.

	adapter	barcode	16S primer
fwd	ATCGCCTACCGTGAC	AAGAAAGTTGTCGGTGTCTTTGTG	AGAGTTTGATCMTGGCTCAG
rev	ATCGCCTACCGTGAC	AAGAAAGTTGTCGGTGTCTTTGTG	CGGTTACCTTGTTACGACTT

samples was approximated from the length distribution and samples were mixed in the proportions listed in Table 15.

TABLE 15. Multiplexing: The sample input mass pooled during multiplexing. The concentration of samples marked with "*" were below the measurement limit and a volume of 0.83 μl was pooled instead.

barcode	run I		run II	
	mass [ng]	ID	mass [ng]	ID
01	<0.1*	NTC	2.1	NTC
02	<0.1*	EC	9.2	EC
03	8.3	ZyDNA	9.0	ZyDNA
04	8.3	ZyCell	9.0	Lu05
05	8.3	St61	9.0	Lu13
06	8.3	St01-1	12.0	Lu18
07	8.3	St02-1	5.7	EC-en
08	8.3	St01-2	3.1	Lu05-en
09	8.3	St03-1	11.7	Lu13-en
10	8.3	St04-1	7.7	Lu18-en
11	8.3	St05-1	9.2	Lu05-en2
12	8.3	St02-2	9.1	Ov85-en
total	83.0	-	96.8	-

The sequencing adapter was added to the pooled barcoded sample amplicons, concluding the library preparation.

4.1.4 Sequencing

A fresh R9.4.1 flow cell (FLO-MIN106D, Oxford Nanopore Technologies, Limited., Oxford, UK) was prepared for each run on a MinION Mk1B (MIN-101B, Oxford Nanopore Technologies, Limited., Oxford, UK) and the pooled library was added to

the flow cell following the manufactures instructions (Oxford Nanopore Technologies, Limited. 2019a).

The sequencing software MinKNOW (Oxford Nanopore Technologies, Limited. 2016) was used to initiate the sequencing run. The folder *00_raw_data* inside the project directory was selected as target. Local basecalling was deactivated. The run duration was set to 48 h for run I and 24 h for run II and the run was initiated. Run II was stopped after 19 h.

4.2 Metabarcoding Analysis Pipeline

4.2.1 A Snakemake Workflow

MeBaPiNa (metabarcoding analysis pipeline for Nanopore datasets) was implemented as a pipeline based on the snakemake workflow management system (version 5.4) (Koster and Rahmann 2012).

The multi complex analysis pipeline (see Figure 2) was divided into smaller steps build around a single process or tool. Each step was integrated in the form of a so-called rule. For each rule, the required input and produced output files were defined, together with the code responsible to the conversion from input to output. The final analysis output was defined in a central rule. As such, upon execution of the *snakemake* command the dependencies were automatically traced back from the final rule to the required input files and all intermediate rules were queued with the correct dependencies. The *--use-conda* parameter was included in the function call, as a conda 4.8.2 (Anaconda, Inc. 2017) based package management was used.

To allow this one line execution, the pipeline was build around a central *Snakemake* file containing the central rule, paths to other rules and the configuration file and other requirements, following the provided template.

4.2.2 Metadata

Two metadata files were provided to MeBaPiNa. A spreadsheet file containing the sample preparation information and a yaml configuration file containing crucial parameters for adapting the pipeline.

All sequenced samples were recorded in the spreadsheet row-wise and the following minimal required columns were extracted upon execution of the pipeline:

- *Sample name*: An ideally unique sample name
- *Run ID*: The unique run identifier produced by the sequencer
- *Flow cell product*: The utilized flow cell product (e.g. *FLO-MIN106*)
- *Sequencing kit*: The utilized sequenting kit (e.g. *SQK-RAB204*)
- *Barcoding kit*: The utilized barcoding kit, if applicable (e.g. *SQK-RAB204*)
- *Barcode*: The barcode assigned to the sample (e.g. *barcode01*)
- *Lambda DCS*: Whether DCS has been added (e.g. *False*)

The following information were specified in the configuration file:

- *experiments*:
  - *project*: Path to the directory containing the *00_raw_data*
  - *meta*: Path to the spreadsheet mentioned above
  - *samples*: List of samples to be analyzed
- *methodologie*: Select central methodologies (one of *kmer*, *align*, *otu*, *asv*)
- *workstation*:
  - *gpu*: Availability of a guppy compatible GPU (either *True* or *False*)
  - *cpu*: Number of available CPU cores (e.g. *42*)
- *filtering*:
  - *q_min*: Minimal average Phred quality per read (e.g. *7*)
  - *len_min*: Minimal sequence length (for reads and references, e.g. *1000*)
  - *len_max*: Maximal sequence length (for reads and references, e.g. *2800*)
  - *min_featurereads*: Minimal number of reads per feature or taxon (e.g. *3*)
  - *min_readidentity*: Read identity threshold (number between *0.0* and *1.0*)
  - *min_confidence*: Confidence threshold in taxonomic classification of lower taxa (number between *0.0* and *1.0*)
- *reference*:
  - *source*: Database source (currently only *silva* is supported)
  - *rank*: Classification rank (one of *species*, *genus*)

The samples specified in the configuration file were used to extract the matching rows from the spreadsheet and define the run identifier and barcodes to include in the analysis. Currently only samples from one run can be analyzed in parallel.

Hence, the run containing the largest overlap with the specified sample set was selected.

4.3 Computational Specifications

Two types of resources were used for execution of the pipeline rules. One desktop PC (Ubuntu 16.04, 2x Nvidia® Quadro® P6000, 2x Intel® Xeon® e5-2650 (12 cores), 270 GB RAM; Dell GmbH, Frankfurt am Main) with guppy compatible GPUs was used for basecalling of the runs. An separate virtual machine (CentOS Linux 7, Intel® Xeon® Gold 6136 (6 cores), 42 GB RAM) on a central server has been used for all subsequent analysis steps. Both devices were connected to a central network accessible storage device (DSM 6.2.2, Intel® Celeron® J3455 (4 Cores), 4 GB RAM; Synology GmbH, Düsseldorf).

4.4 Quality Control

4.4.1 Basecalling, Demultiplexing and Trimming

The ONT basecaller guppy (Oxford Nanopore Technologies, Limited. 2018a) was used for basecalling of the sequencing runs selected through their run identifier. The version and parameters of the invoked *guppy_basecaller* function are listed in Table 16.

The demultiplexed reads were separated into a *pass* and *fail* directory depending on their average quality and the specified threshold and detected calibration strands were separated out, if their inclusion was specified in the spreadsheet (see Section 4.2.2).

The resulting barcode specific folders containing the fastq files were individually passed to qcat 1.1.0 (Oxford Nanopore Technologies, Limited. 2018b), to perform a second round of demultiplexing and to trim off sequencing adapter, barcode region and primer sequences from both read ends and split reads if a barcode was detected in the middle of the read. Therefore, the *--trim* and *--detect-middle* parameters were passed to the *qcat* function together with setting the barcode mapping score to 70 (*--min-score 70*) and *--kit* setting to *RAB204*.

TABLE 16. Parameters for guppy basecaller: Version, parameters and variables used for basecalling of both sequencing runs. "-" indicate the absence of an parameter without variable in the function call, "+" indicates its inclusion.

parameter	**run** I	**run** II
version	3.2.4	3.4.1
--flowcell	FLO-MIN106	FLO-MIN106
--kit	SQK-RAD204	SQK-RAD204
--barcode_kits	SQK-RAB204	SQK-RAB204
--calib_detect	-	-
--qscore_filtering	+	+
--min_qscore	0	0
--device	cuda:all:100%	cuda:all:100%
--gpu_runners_per_device	4	6
--chunks_per_runner	512	1536
--chunk_size	1000	1000
--chunks_per_caller	10000	10000
--num_barcode_threads	4	8
--num_callers	12	8
--fast5_out	+	+

4.4.2 Filtering

Filtering was applied through NanoFilt 2.6.0 (De Coster et al. 2018). The barcode specific fastq file containing the trimmed reads and the *sequencing_summary.txt* file, produced by the basecaller, were specified as input for the *NanoFilt* function. The filtering variables, specified in the configuration file, were passed to *NanoFilt* to exclude reads below 1000 bp (*--length 1000*) and above 2800 bp (*--maxlength 2800*) as well as reads below a Phred quality score of 7 (*--quality 7*).

4.4.3 Visualization and Statistics

Functions from the NanoPack (De Coster et al. 2018) tool set were used to produce visualizations for the quality control steps. The read length and quality distribution scatter plots before and after filtering, the spatial flow cell throughput visualization as well as the time course of active pores and throughput were produced by *NanoPlot* 1.28.0. The *sequencing_summary.txt* was specified as input for basecalled reads via the *--summary* parameter and filtered fastq files were used for filtered reads

via the *--fastq_rich* parameter. Other parameters were set as *--maxlength 2800*, *--drop_outliers*, *--plots kde hex dot*, *--colormap viridis*, *--color black* and *--downsample 100000* to change the plot aesthetics. The plots for sequencing throughput, read length and quality distribution of the demultiplexed reads before and after filtering were produced by *NanoComp* 1.9.2 with parameters *--maxlength 2800*, *--barcoded*, *--plot violin* and the same input as *NanoPlot*. The positional-quality-score plot of all basecalled reads was produced by *nanoQC* 0.9.2 with the parameter *--minlen 240*.

Statistics on the number of reads, bases, read length and Phred quality scores were extracted from the *report.md*, produced by the sequencer for the raw reads and *NanoStats.txt* produced by *NanoComp* for the processed reads. Sums and averages were calculated using custom command line scripts.

4.5 Reference Database

The aligned reference sequence file

SILVA_138_SSURef_NR99_tax_silva_full_align_trunc.fasta.gz,

and taxonomic association files

taxmap_slv_ssu_ref_nr_138.txt.gz

and

tax_slv_ssu_138.txt.gz,

of release 138 were downloaded from the FTP server of the SILVA project (Yilmaz et al. 2014), the checksum was validated and the files decompressed.

The reference sequences were converted to gaped DNA sequences by the *convert_rna_to_dna.py* function provided by Robeson (2020) including the *--convert_to_gap* option. The primer sequences from Section 4.1.3 were aligned to the first 500 sequences converted reference sequences using mafft 7.310 (Katoh and Standley 2013), by calling the equally named function with the *--addfragments* parameter. Start and end position of the primer alignment were extracted from the returned mapping file using awk (Free Software Foundation, Inc. 2019) and supplied to *extract_alignment_region.py* (Robeson 2020) via the *--start_position* and *--end_position* parameter. The resulting extracted amplicon

region was subsequently degaped using the *degap_fasta.py* script (Robeson 2020). Sequences below or above the length thresholds provided in the configuration file from Section 4.2.2 were filtered out by vsearch 2.7.0 (Rognes et al. 2016). The length thresholds were passed to the *vsearch* function via the *--fastq_minlen* and *--fastq_maxlen* parameters together with the *--fastx_filter* flag. The vsearch tool was also used to remove duplicate sequences by applying *--derep_fulllength* on the filtered sequences.

A custom python 3.7.6 script (Python Software Foundation 2020) was used to filter the taxonomic association files, include the species rank assignments and produce the output files required for the generation of the method specific reference database files. Briefly, the taxonomic association files were loaded into pandas 1.0.1 (McKinney 2011) data frames and merged on the taxonomic IDs. The sequence specific names contained in the sequence description line (originated from the RefSeq database (O'Leary et al. 2016)) were used as species names and undesirable characters (other than alphanumerical characters or one of -/.[]()\) were removed before the names were reduced to the two first words, to exclude additional strain specifications. The taxonomic paths were split and the species names were set to position seven. Generated gaps were filled with dummy taxa carrying the prefix "_unknown" followed by the rank and were assigned the rank "no rank" to prevent assignment of reads to them. New taxonomic identifiers were created for all new taxa while ensuring the existing SILVA identifiers to be preserved. The data was converted into the file formats and types required by the downstream analysis tools and exported.

A second set of output files was created without the species rank taxonomy which can be selected through the parameter in the configuration file from Section 4.2.2.

4.6 Feature Extraction

4.6.1 Operation Taxonomic Unit Picking

The tools used to performe the feature extraction via OTU picking were implemented in the vsearch 2.7.0 (Rognes et al. 2016) integration of the QIIME2 pipeline (version 2020.2.0) by Bolyen et al. (2019) and other tools of this pipeline.

Dereplication and OTU Picking

The quality controlled reads from Section 4.4.2 were imported for each sample individually using *qiime tools import*. The subsequent dereplication was executed as *qiime vsearch dereplicate-sequences*.

The OTU picking in form of an open-reference clustering was also performed by vsearch using the function *qiime vsearch cluster-features-open-reference* with the dereplicated sequencing reads as input. Further, the set of filtered reference sequences from Section 4.5 was imported via *qiime tools import* and passed to the clustering function as reference. Upon completion, a set of sequences, one per feature, and a table with the number of reads assigned to each feature was reported.

Chimera Detection and Filtering

The subsequent chimera detection has been performed by the vsearch integration of QIIME2 as well. Chimeras were detected in the feature sequences by passing the OTU picking output to *qiime vsearch uchime-denovo* and obtaining separate files for feature sequences predicted to be chimeric and non-chimeric features.

The set of chimeric sequences were excluded and low abundance features were filtered by utilizing QIIME2s *qiime feature-table filter-features* with *--p-min-frequency 3* and *--p-exclude-ids* on the feature count table and extracting the associated sequences by invoking *qiime feature-table filter-seqs* with *--p-no-exclude-ids* on the feature sequences.

Statistics

The number of de-novo features was extracted from the verbose output of *qiime vsearch uchime-denovo* and the number of reads and features before and after chimera removal and filtering were extracted from the feature tables using awk (Free Software Foundation, Inc. 2019), while calculating the median at the same time.

4.6.2 Amplicon Sequence Variants

The DADA2 1.12.1 package (Callahan et al. 2016) for R 3.6.1 (R Core Team 2019) was used to investigate the possibility of error correction through detection

of amplicon sequence variants.

Reads were dereplicated using the *derepFastq* function. The error model was inferred by the *learnErrors* function with the paramerer *multithread=TRUE* and visualized through *plotErrors* with *nominalQ=TRUE*. No further functions were implemented as the inference of the error model, required for further processing, failed (see Section 2.4.5).

4.7 Taxonomic Classification

4.7.1 Naive Bayes

The scikit-learn 0.22.1 (Pedregosa et al. 2011) integrated in the QIIME2 pipeline (version 2020.2.0) (Bolyen et al. 2019) was implemented for classification of extracted features.

Training of the classifier on the filtered reference sequences and modified taxonomic associations from Section 4.5, imported via *qiime tools import*, was implemented by calling *qiime feature-classifier fit-classifier-naive-bayes*.

The trained classifier was required for the taxonomic classification of the feature clusters described in Section 4.6.1 via *qiime feature-classifier classify-sklearn*.

4.7.2 K-mer Mapping

The k-mer mapping was implemented via the Kraken 2 2.0.8_beta tool set (Wood et al. 2019). The subsequent reestimation of abundance was based on functions from the Bracken 2.5 tool set (Lu et al. 2017).

Reference Database

A reference database was created from the filtered reference sequences and modified taxonomic associations produced as described in Section 4.5, by executing *kraken2-build* with *--kmer-len 35* and *--build*, followed by *bracken-build* with *-k 35* and *-l 1451*. The resulting reference database for a k-mer length of 35 and a

expected read length of 1451 (observed median read length in run I) was used for the analysis of both sequencing runs.

K-mer Mapping, Reestimation

The created database was used to perform k-mer mapping of the quality controlled reads from Section 4.4.2 via the *kraken2* function with a defined confidence threshold (*--confidence 0.0*) and the created database as reference.

The taxonomic classification abundance in the created report file was reestimated for species ranks using the *bracken* function with *-r 1451* and *-l S*.

The report with the reestimated abundances was converted to an table containing the full taxa path by an custom python 3.7.6 script (Python Software Foundation 2020) using pandas 1.0.1 (McKinney 2011).

OTU Picked Features

The k-mer mapping of the clustered features was implemented as an alternative to the naïve Bayes approach from Section 4.7.1. The information from the filtered feature table was used in awk (Free Software Foundation, Inc. 2019) to replicate the representative sequence of each feature times the number of reads assigned to the feature, to get a simulation of the full read set. The produced read file was used for k-mer mapping as described above.

4.7.3 Full-Length Alignment

Indexing and Alignemnt

The reference sequences were indexed using minimap2 2.17 (Li 2018) by invoking the equally named function with the *-d* parameter and setting the use case to *-x map-ont*.

This index was then used to align quality controlled reads from Section 4.4 to the reference sequences by running *minimap2* with the same use case parameter (*-x map-ont*). The sam format was specified as output (triggered by including parameter *-a* to the function call) to get position accurate CIGAR alignment used for statistics

and filtering. Further, secondary alignments were only allowed when resulting in the same highest score (*-p 1*) and only one additional alignment, besides the primary, was reported (*-N 1*).

Filtering, Collapsing and Conversion

Chimeric, supplementary and reads with multiple alignments were filtered out using samtools 1.9 (Li et al. 2009), by calling *samtools view* with the parameters *-F 2048* and *-q 1*.

The filtering was continued with a custom python 3.7.6 script (Python Software Foundation 2020) including pandas 1.0.1 (McKinney 2011). The prefiltered alignment sam file was imported and the length of the aligned segment was calculated from the CIGAR string, excluding clipped bases. Reads with an alignment segment length below the threshold of 1000 bp or a gap-compressed per-base sequence divergence (reported by the *de* flag of minimap2) above 0.1 were filtered out.

The taxonomic assignment reference from Section 4.4.2 was loaded and reads were collapsed per reference sequence ID (*RNAME*) and reads per reference sequence were collapsed per taxonomic ID via their association. The taxonomic abundances were saved in a table containing the full taxa path (same format as produced after reestimation of abundance).

4.7.4 Visualization and Statistics

The density plot of the alignment identity and Phred score was produced by pycoQC 2.5.0.17 (Leger and Leonardi 2019). The *pycoQC* function was called with a custom config file (via *--config*) to exclude the coverage plot, as the number of reference contigs was to high. Further, the quality filtering was deactivated (*--min_pass_qual 0*) and the data was downsampled (*--sample 100000*). The plot of interest was manually exported from the html report.

Krona 2.7.1 (Ondov et al. 2011) was used to produce the multi-layer pie-charts for visualization of the taxonomic composition. The *ktImportTaxonomy* function was used for the Kraken 2 output files with the parameters *-t 3*, *-d 7* and -i to create the desired output and *ktImportText* with the parameter *-n "Root"* was used to plot

taxonomic abundance tables (containing the full taxonomic path) from the full-length alignment and reestimation of abundance.

The information about the classified taxa, number of assigned reads and median reads were extracted from the report file of Kraken 2 and Bracken and the converted taxonomic abundance table from the full-length alignment using awk (Free Software Foundation, Inc. 2019).

The statistical properties of the communities were calculated with the vegan 2.5_6 package (Oksanen et al. 2019) under R 3.6.1 (R Core Team 2019). Richness R was calculated using the *specnumber* function and is defined as the number of classified species rank taxa. Shannon entropy H for the observed richness R was determined by the *diversity* function and the parameter *index = "shannon"* and *base = exp(1)*. It was calculated as:

$$H = -\sum_{i=1}^{R} p_i * ln(p_i)$$

where p_i is the proportion reads assigned to species i. The Simpson index λ was calculated by the same function invoked with *index = "simpson"* instead and is defined as:

$$\lambda = 1 - \sum_{i=1}^{R} p_i^2$$

Evenness J was calculated as the quotient of richness and the logarithm of the Shannon entropy (Hill 1973; Oksanen et al. 2019):

$$J = H/log(R)$$

The relative abundance and relative deviation from the reference of the mock community have been created in a custom R 3.6.1 (R Core Team 2019) script build around ggplot 3.1.1 (Wickham 2016), based on the taxonomic read assignments extracted as described above.

4.8 Availability

The MeBaPiNa code used in this work is available at GitHub (https://github.com/Marc-Ruebsam/MeBaPiNa). For the availability of sample sequencing data, please

contact PD Dr. med. Niels HALAMA (niels.halama (at) nct-heidelberg.de).

References

Anaconda, Inc. (2017). Conda. https://conda.io/en/latest/. Revision a660a85d.

Balvočiūtė, M. and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics *18*.

Benítez-Páez, A., Portune, K. J. and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. GigaScience *5*, 4.

Benítez-Páez, A. and Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. GigaScience *6*, gix043.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodriguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y. X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vazquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R.

and Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857.

Bouchet, V., Huot, H. and Goldstein, R. (2008). Molecular Genetic Basis of Ribotyping. Clinical Microbiology Reviews *21*, 262–273.

Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T. P., Puetz, L., Rasmussen, M., Zervas, A. and Hansen, L. H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience *9*.

Byrne, A., Cole, C., Volden, R. and Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences *374*, 20190097.

Callahan, B. (2019). Nanopore consensus sequencing and DADA2. https://github.com/benjjneb/dada2/issues/759.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods *13*, 581–583.

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., Dworkin, J. P., Lupisella, M. L., Smith, D. J., Botkin, D. J., Stephenson, T. A., Juul, S., Turner, D. J., Izquierdo, F., Federman, S., Stryke, D., Somasekar, S., Alexander, N., Yu, G., Mason, C. E. and Burton, A. S. (2017). Nanopore DNA Sequencing and Genome Assembly on the International Space Station. Scientific Reports *7*.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. and Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–642.

Curren, E., Yoshida, T., Kuwahara, V. S. and Leong, S. C. Y. (2019). Rapid profiling of tropical marine cyanobacterial communities. Regional Studies in Marine Science *25*, 100485.

Cuscó, A., Catozzi, C., Viñes, J., Sanchez, A. and Francino, O. (2019). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA

gene and the 16S-ITS-23S of the rrn operon. F1000Research *7*, 1755.

Cuscó, A., Viñes, J., D'Andreano, S., Riva, F., Casellas, J., Sánchez, A. and Francino, O. (2017). Using MinION to characterize dog skin microbiota through full-length 16S rRNA gene sequencing approach. bioRxiv *1*.

Daubin, V., Moran, N. A. and Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science *301*, 829–832.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669.

de Lannoy, C., de Ridder, D. and Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. F1000Res *6*, 1083.

DeAngelis, M. M., Wang, D. G. and Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Research *23*, 4742–4743.

Derakhshani, H., Tun, H. M. and Khafipour, E. (2015). An extended single-index multiplexed 16S rRNA sequencing for microbial community analysis on MiSeq illumina platforms. Journal of Basic Microbiology *56*, 321–326.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods *10*, 996–998.

Edgar, R. C. (2016a). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv *1*.

Edgar, R. C. (2016b). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv *1*.

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ *5*, e3889.

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ *6*, e4652.

Edwards, A., Debbonaire, A. R., Nicholls, S. M., Rassner, S. M., Sattler, B., Cook, J. M., Davy, T., Soares, A., Mur, L. A. and Hodson, A. J. (2016). In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. bioRxiv *1*.

Enaud, R., Prevel, R., Ciarlo, E., Beaufils, F., Wieërs, G., Guery, B. and Delhaes, L. (2020). The Gut-Lung Axis in Health and Respiratory Diseases: A Place for Inter-Organ and Inter-Kingdom Crosstalks. Frontiers in Cellular and Infection Microbiology *10*.

Falony, G., Vandeputte, D., Caenepeel, C., Vieira-Silva, S., Daryoush, T., Vermeire, S. and Raes, J. (2019). The human microbiome in health and disease: hype or hope. Acta Clinica Belgica *74*, 53–64.

Feibelman, T., Bayman, P. and Cibula, W. G. (1994). Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. Mycological Research *98*, 614–618.

Free Software Foundation, Inc. (2019). GAWK: Effective AWK Programming: A User's Guide for GNU Awk, for the 5.0.0 (or later) version of the GNU implementation of AWK.

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., de Macedo, M. P., Cotechini, T., Kumar, T., Chen, W. S., Reddy, S. M., Sloane, R. S., Galloway-Pena, J., Jiang, H., Chen, P. L., Shpall, E. J., Rezvani, K., Alousi, A. M., Chemaly, R. F., Shelburne, S., Vence, L. M., Okhuysen, P. C., Jensen, V. B., Swennes, A. G., McAllister, F., Sanchez, E. M. R., Zhang, Y., Chatelier, E. L., Zitvogel, L., Pons, N., Austin-Breneman, J. L., Haydu, L. E., Burton, E. M., Gardner, J. M., Sirmans, E., Hu, J., Lazar, A. J., Tsujikawa, T., Diab, A., Tawbi, H., Glitza, I. C., Hwu, W. J., Patel, S. P., Woodman, S. E., Amaria, R. N., Davies, M. A., Gershenwald, J. E., Hwu, P., Lee, J. E., Zhang, J., Coussens, L. M., Cooper, Z. A., Futreal, P. A., Daniel, C. R., Ajami, N. J., Petrosino, J. F., Tetzlaff, M. T., Sharma, P., Allison, J. P., Jenq, R. R. and Wargo, J. A. (2017). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science *359*, 97–103.

Gorzelak, M. A., Gill, S. K., Tasnim, N., Ahmadi-Vand, Z., Jay, M. and Gibson, D. L. (2015). Methods for improving human gut microbiome data by reducing variability

through sample processing and storage of stool. PloS one *10*, e0134802.

Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. and Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. Nature Medicine *25*, 377–388.

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology *54*, 427–432.

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., Martellaro, C., Falzarano, D., Marzi, A., Squires, R. B., Wollenberg, K. R., de Wit, E., Prescott, J., Safronetz, D., van Doremalen, N., Bushmaker, T., Feldmann, F., McNally, K., Bolay, F. K., Fields, B., Sealy, T., Rayfield, M., Nichol, S. T., Zoon, K. C., Massaquoi, M., Munster, V. J. and Feldmann, H. (2016). Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. Emerging Infectious Diseases *22*.

Hugerth, L. W. and Andersson, A. F. (2017). Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Front Microbiol *8*, 1561.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Brooks, J., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S., Chen, I. M., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Dunne, W. M., Durkin, A., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. A., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M.,

Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haake, S. K., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C. C., Lozupone, C. A., Lunsford, R., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y. H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Sathirapongsasuti, J., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Methe, B. A., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K. and White, O. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J. and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Jain, M., Olsen, H. E., Paten, B. and Akeson, M. (2016). The Oxford Nanopore

MinION: delivery of nanopore sequencing to the genomics community. Genome Biology *17*.

Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., Malla, S., Leggett, R. M., Wallerman, O., Jansen, H. J., Zalunin, V., Birney, E., Brown, B. L., Snutch, T. P. and Olsen, H. E. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Res *6*, 760.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E. and Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature Communications *10*.

Kai, S., Matsuo, Y., Nakagawa, S., Kryukov, K., Matsukawa, S., Tanaka, H., Iwai, T., Imanishi, T. and Hirota, K. (2019). Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION nanopore sequencer. FEBS Open Bio *9*, 548–557.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780.

Kerkhof, L. J., Dillon, K. P., Haggblom, M. M. and McGuinness, L. R. (2017). Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome *5*, 116.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glöckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research *41*, e1.

Kono, N. and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Development, Growth & Differentiation *61*, 316–326.

Koster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Leger, A. and Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. The Journal of Open Source Software *4*, 1236.

Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A. and Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. GigaScience *5*, 34.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and and, R. D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*.

Lu, J., Breitwieser, F. P., Thielen, P. and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science *3*, e104.

Ma, X., Stachler, E. and Bibby, K. (2017). Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization. bioRxiv *1*.

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A. and Abd_Allah, E. F. (2018). Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. Frontiers in immunology *9*.

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J. and Gajewski, T. F. (2018). The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. Science *359*, 104–108.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P. (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal *6*, 610–618.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing *14*.

McOrist, A. L., Jackson, M. and Bird, A. R. (2002). A comparison of five methods for extraction of bacterial DNA from human faecal samples. J. Microbiol. Methods *50*, 131–139.

Minot, S. S., Krumm, N. and Greenfield, N. B. (2015). One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. bioRxiv *1*.

Mitsuhashi, S., Kryukov, K., Nakagawa, S., Takeuchi, J. S., Shiraishi, Y., Asano, K. and Imanishi, T. (2017). A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. Scientific Reports *7*, 5657.

Moffatt, M. F. and Cookson, W. O. (2017). The lung microbiome in health and disease. Clinical Medicine *17*, 525–529.

Moon, J., Kim, N., Lee, H. S., Shin, H. R., Lee, S. T., Jung, K. H., Park, K. I., Lee, S. K. and Chu, K. (2017). Campylobacter fetus meningitis confirmed by a 16S rRNA gene analysis using the MinION nanopore sequencer, South Korea, 2016. Emerging Microbes & Infections *6*, 1–3.

Myer, P. R., Kim, M., Freetly, H. C. and Smith, T. P. L. (2016). Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. J. Microbiol. Methods *127*, 132–140.

O'Dwyer, D. N., Dickson, R. P. and Moore, B. B. (2016). The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease. The Journal of Immunology *196*, 4839–4847.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. (2019). vegan: Community Ecology Package. R package version 2.5-6.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch,

C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–745.

Ondov, B. D., Bergman, N. H. and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics *12*.

Oxford Nanopore Technologies, Limited. (2016). MinKNOW. Oxford, UK. https://community.nanoporetech.com/technical_documents/minknow-tech-doc/v/mitd_5000_v1_revy_16may2016.

Oxford Nanopore Technologies, Limited. (2018a). Guppy. Oxford, UK. https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revp_14dec2018, Access restricted.

Oxford Nanopore Technologies, Limited. (2018b). qcat. GitHub repository, https://github.com/nanoporetech/qcat/tree/e27a1127485b75b76e2f5dc23359010ed05c5d2b.

Oxford Nanopore Technologies, Limited. (2019a). 16S Barcoding Kit (SQK-RAB204). Oxford, UK. https://community.nanoporetech.com/protocols/16S-barcoding-sequencing/v/RAB_9053_v1_revL_14Aug2019, Access restricted.

Oxford Nanopore Technologies, Limited. (2019b). Product Comparison. https://nanoporetech.com/products/comparison.

Oxford Nanopore Technologies, Limited. (2019c). EPI2ME 16S workflow: real-time identification of bacteria and archaea. Oxford, UK.

Payne, A., Holmes, N., Rakyan, V. and Loose, M. (2018). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics *35*, 2193–2198.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Pflughoeft, K. J. and Versalovic, J. (2012). Human Microbiome in Health and Disease. Annual Review of Pathology: Mechanisms of Disease *7*, 99–122.

Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Appl. Environ. Microbiol. *84*.

PROMISE trial consortium (2018). Prädiktive immunologische signaturen bei lungenkrebs. clinical trial currently ongoing.

Python Software Foundation (2020). The Python Language Reference.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. version 3.7.

Rang, F. J., Kloosterman, W. P. and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. *19*, 90.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G., Gasbarrini, A. and Mele, M. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. Microorganisms *7*, 14.

Robeson, M. (2020). make_SILVA_db. GitHub repository, https://github.com/mikerobeson/make_SILVA_db/tree/0823e8771fa119c83c4be2e545adbd44e4e5a48e.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

Routy, B., Chatelier, E. L., Derosa, L., Duong, C. P. M., Alou, M. T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragón, L., Jacquelot, N., Qu, B., Ferrere, G., Clémenson, C., Mezquita, L., Masip, J. R., Naltet, C., Brosseau, S., Kaderbhai, C., Richard, C., Rizvi, H., Levenez, F., Galleron, N., Quinquis, B., Pons, N., Ryffel, B., Minard-Colin, V., Gonin, P., Soria, J.-C., Deutsch, E., Loriot, Y., Ghiringhelli, F., Zalcman, G., Goldwasser, F.,

Escudier, B., Hellmann, M. D., Eggermont, A., Raoult, D., Albiges, L., Kroemer, G. and Zitvogel, L. (2017). Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science *359*, 91–97.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J. and Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87.

Savage, D. C. (1977). Microbial Ecology of the Gastrointestinal Tract. Annual Review of Microbiology *31*, 107–133.

Schloss, P. D. (2009). A High-Throughput DNA Sequence Aligner for Microbial Ecology Studies. PLoS ONE *4*, e8230.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Sender, R., Fuchs, S. and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biology *14*, e1002533.

Shin, H., Lee, E., Shin, J., Ko, S. R., Oh, H. S., Ahn, C. Y., Oh, H. M., Cho, B. K. and Cho, S. (2018). Elucidation of the bacterial communities associated with the harmful microalgae Alexandrium tamarense and Cochlodinium polykrikoides using nanopore sequencing. Scientific Reports *8*, 5323.

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H. and Cho, B.-K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Scientific Reports *6*, 29681.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. and Tolstoy, I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. *42*, D553–559.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends in Genetics *34*, 666–681.

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J. and Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome *3*, 26.

Wang, B., Yao, M., Lv, L., Ling, Z. and Li, L. (2017). The Human Microbiota in Health and Disease. Engineering *3*, 71–82.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology *73*, 5261–5267.

Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S. and Qian, P.-Y. (2014). Optimal Eukaryotic 18S and Universal 16S/18S Ribosomal RNA Primers and Their Application in a Study of Symbiosis. PLoS ONE *9*, e90053.

Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J. and Knight, R. (2014). Tracking down the sources of experimental contamination in microbiome studies. Genome Biology *15*.

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., Angenent, L. T., Knight, R. and Ley, R. E. (2011). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. The ISME Journal *6*, 94–103.

Wick, R. R., Judd, L. M. and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. *20*, 129.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. *74*, 5088–5090.

Wood, D. E., Lu, J. and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology *20*.

Wright, E. S. and Vetsigian, K. H. (2016). Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics *17*.

Yang, B., Wang, Y. and Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics *17*, 135.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. and Glockner, F. O. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. *42*, D643–648.

Zhou, B., Sun, C., Huang, J., Xia, M., Guo, E., Li, N., Lu, H., Shan, W., Wu, Y., Li, Y., Xu, X., Weng, D., Meng, L., Hu, J., Gao, Q., Ma, D. and Chen, G. (2019). The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients. Scientific Reports *9*.

Supplement

S.I Supplementary Tables

TABLE S1. Read statistics per sample: A list of read statistics for the demultiplexed samples (run I top, run II bottom, see Table 1 for reference). Shown are the total number of reads, total bases, median length and median quality after demultiplexing (dem), after trimming and filtering (flt). Unassigned reads are presented as ID "-".

ID	reads [$*10^6$]		bases [Gb]		length [kb]		quality	
	dem	flt	dem	flt	dem	flt	dem	flt
NTC	<0.1	<0.1	<0.1	<0.1	0.2	1.5	7.7	11.7
EC	<0.1	<0.1	<0.1	<0.1	0.2	1.5	7.5	12.2
ZyDNA	0.8	0.7	1.2	1.0	1.6	1.5	10.5	11.8
ZyCell	1.8	1.5	2.8	2.2	1.6	1.5	10.5	11.8
St61	0.8	0.7	1.3	1.0	1.6	1.4	10.5	11.8
St01-1	1.1	0.9	1.6	1.2	1.6	1.4	10.4	11.7
St02-1	0.8	0.7	1.2	1.0	1.6	1.4	10.6	11.8
St01-2	0.6	0.5	1.0	0.8	1.6	1.4	10.5	11.8
St03-1	1.6	1.3	2.4	1.9	1.6	1.4	10.5	11.8
St04-1	1.1	1.0	1.8	1.4	1.6	1.4	10.5	11.7
St05-1	0.8	0.7	1.3	1.0	1.6	1.4	10.5	11.7
St02-2	0.7	0.6	1.1	0.9	1.6	1.4	10.4	11.7
-	1.8	-	2.6	-	1.6	-	5.3	-
NTC	0.1	0.1	0.1	0.1	1.6	1.4	10.5	11.7
EC	0.4	0.3	0.6	0.4	1.5	1.4	10.2	11.5
ZyDNA	1.0	0.7	1.3	1.0	1.6	1.5	10.3	11.6
Lu05	0.6	0.2	0.5	0.3	0.4	1.4	9.7	11.6
Lu13	0.4	0.2	0.4	0.3	1.4	1.4	10.0	11.7
Lu18	2.2	0.2	1.1	0.2	0.4	1.4	9.5	11.4
EC-en	0.2	0.1	0.2	0.2	1.6	1.4	10.5	11.8
Lu05-en	0.3	0.1	0.2	0.1	0.4	1.4	9.5	11.5
Lu13-en	0.4	0.1	0.3	0.1	0.4	1.5	9.7	11.7
Lu18-en	0.2	0.2	0.3	0.2	1.6	1.4	10.4	11.8
Lu05-en2	0.4	0.2	0.4	0.3	1.6	1.5	10.2	11.7
Ov85-en	1.2	0.9	1.7	1.3	1.6	1.4	10.3	11.6
-	1.3	-	1.2	-	0.6	-	5.8	-

TABLE S2. Alignment error profile: List of insertion (ins), deletion (deletion), mismatch (mis) and combined (all) error rates of all reads as well as their median alignment identity (ali) and the expected identity (exp) calculated from the Phred quality scores in Table S1 of all reads per sample (run I top, run II bottom, see Table 1 for reference).

ID	error rate [%]				identity [%]	
	all	ins	del	mis	ali	exp
NTC	7.8	2.3	2.8	2.7	92.5	93.2
EC	8.3	2.3	3.4	2.6	92.6	94.0
ZyDNA	9.0	2.9	3.1	3.0	91.8	93.4
ZyCell	9.1	2.9	3.1	3.1	91.6	93.4
St61	9.0	2.8	3.1	3.2	91.7	93.4
St01-1	9.5	2.8	3.3	3.4	91.2	93.2
St02-1	8.9	2.8	3.1	3.1	91.8	93.4
St01-2	9.1	2.8	3.1	3.2	91.6	93.4
St03-1	9.1	2.9	3.0	3.2	91.6	93.4
St04-1	9.2	2.8	3.1	3.2	91.6	93.2
St05-1	9.3	2.8	3.2	3.3	91.5	93.2
St02-2	9.3	2.8	3.2	3.3	91.5	93.2
NTC	8.8	2.3	3.3	3.2	92.0	93.2
EC	9.6	2.5	3.7	3.5	91.3	92.9
ZyDNA	11.2	2.6	3.7	4.8	89.2	93.1
Lu05	9.7	2.6	3.4	3.7	91.0	93.1
Lu13	9.0	2.3	3.3	3.3	91.8	93.2
Lu18	9.8	2.5	3.8	3.5	91.0	92.8
EC-en	8.8	2.3	3.3	3.2	92.0	93.4
Lu05-en	9.5	2.5	3.7	3.3	91.3	92.9
Lu13-en	8.7	2.5	3.1	3.1	92.0	93.2
Lu18-en	8.8	2.2	3.2	3.3	92.0	93.4
Lu05-en2	8.8	2.4	3.2	3.2	92.0	93.2
Ov85-en	9.5	2.6	3.4	3.5	91.2	93.1

TABLE S3. Taxonomic coverage – OTU picking, unfiltered: The number of total (all) and de novo (de novo) features produced by open reference clustering during OUT picking and reads associated with the features as well as the median reads per feature of each sample (run I top, run II bottom, see Table 1 for reference).

ID	reads $[*10^6]$	features $[*10^3]$ all	features $[*10^3]$ de novo	median
NTC	<0.1	<0.1	<0.1	1
EC	<0.1	<0.1	<0.1	1
ZyDNA	0.7	162.7	160.6	1
ZyCell	1.5	386.0	383.4	1
St61	0.7	199.9	196.5	1
St01-1	0.9	251.9	247.0	1
St02-1	0.7	177.4	175.1	1
St01-2	0.5	141.7	139.9	1
St03-1	1.3	326.7	322.6	1
St04-1	1.0	255.9	252.9	1
St05-1	0.7	199.6	197.0	1
St02-2	0.6	170.7	168.3	1
NTC	0.1	16.3	16.1	1
EC	0.3	73.5	73.1	1
ZyDNA	0.7	242.6	239.6	1
Lu05	0.2	52.2	51.1	1
Lu13	0.2	45.8	45.6	1
Lu18	0.2	49.9	49.3	1
EC-en	0.1	33.4	33.3	1
Lu05-en	0.1	17.1	16.8	1
Lu13-en	0.1	16.4	16.1	1
Lu18-en	0.2	33.0	32.8	1
Lu05-en2	0.2	52.2	52.1	1
Ov85-en	0.9	232.4	230.8	1

S.II Supplementary Figures

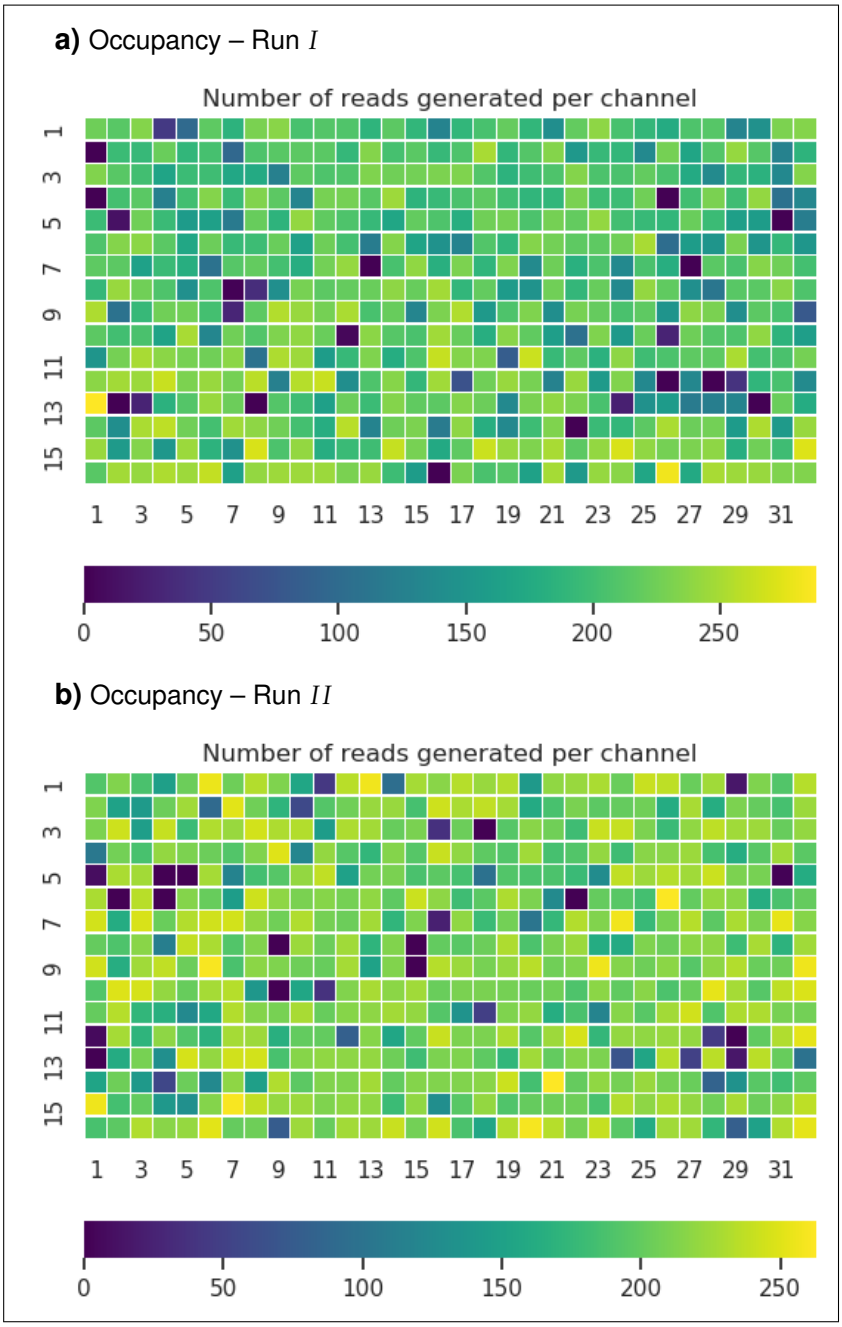

FIGURE S1. Spatial flow cell throughput: Spatial representation of the 512 channels on the flow cell chip of run I (Subfigure **a)**) and run II (Subfigure **b)**). The color corresponds to the cumulative throughput of each channel over the course of the experiment.

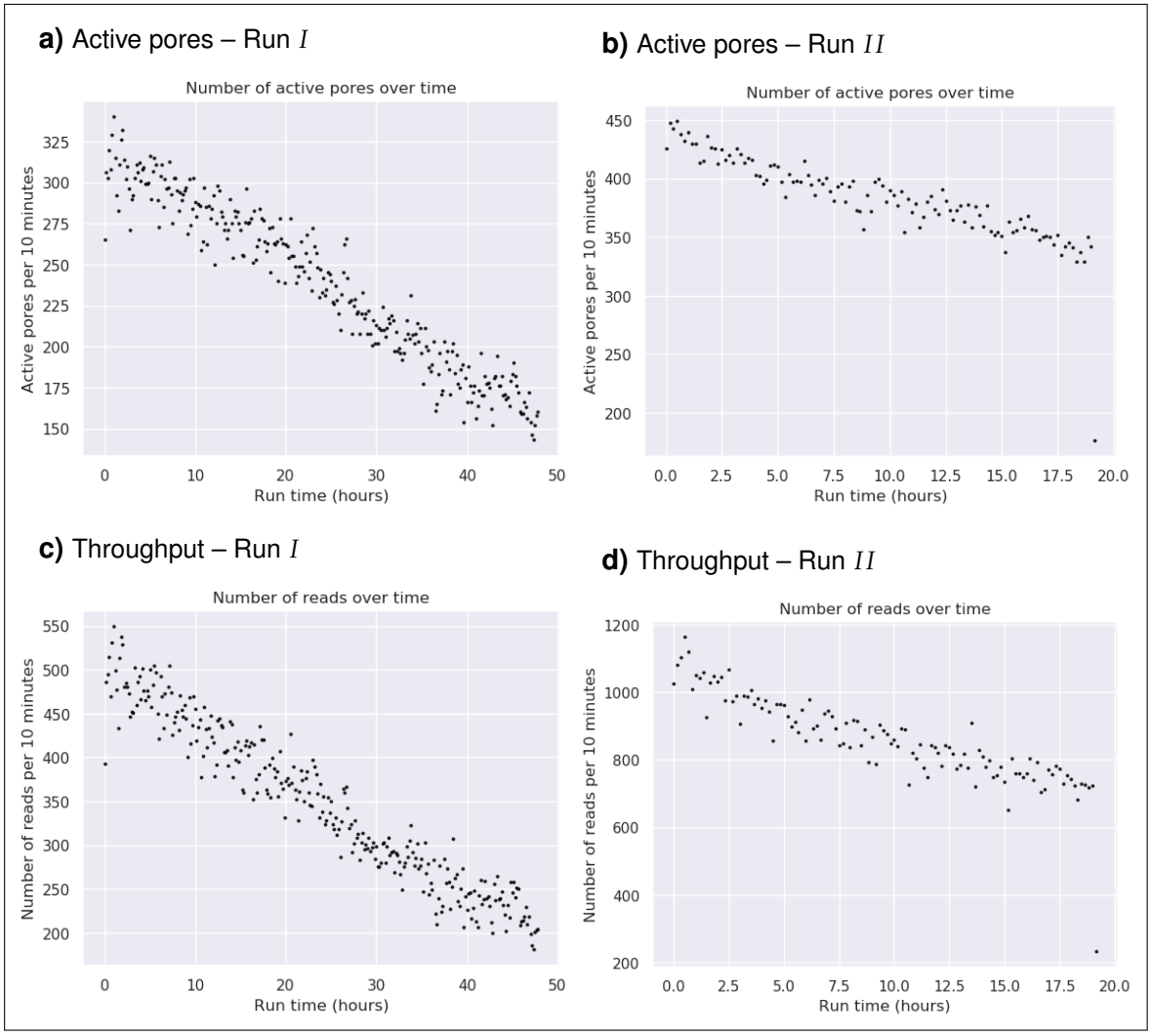

FIGURE S2. Course of active pores and throughput: Shown are the number of active pores (Subfigure **a)** and **b)**) and read throughput (Subfigure **c)** and **d)**) over the time course of the experiment binned into 10 minute intervals for run I (Subfigure **a)** and **c)**) and run II (Subfigure **b)** and **d)**).

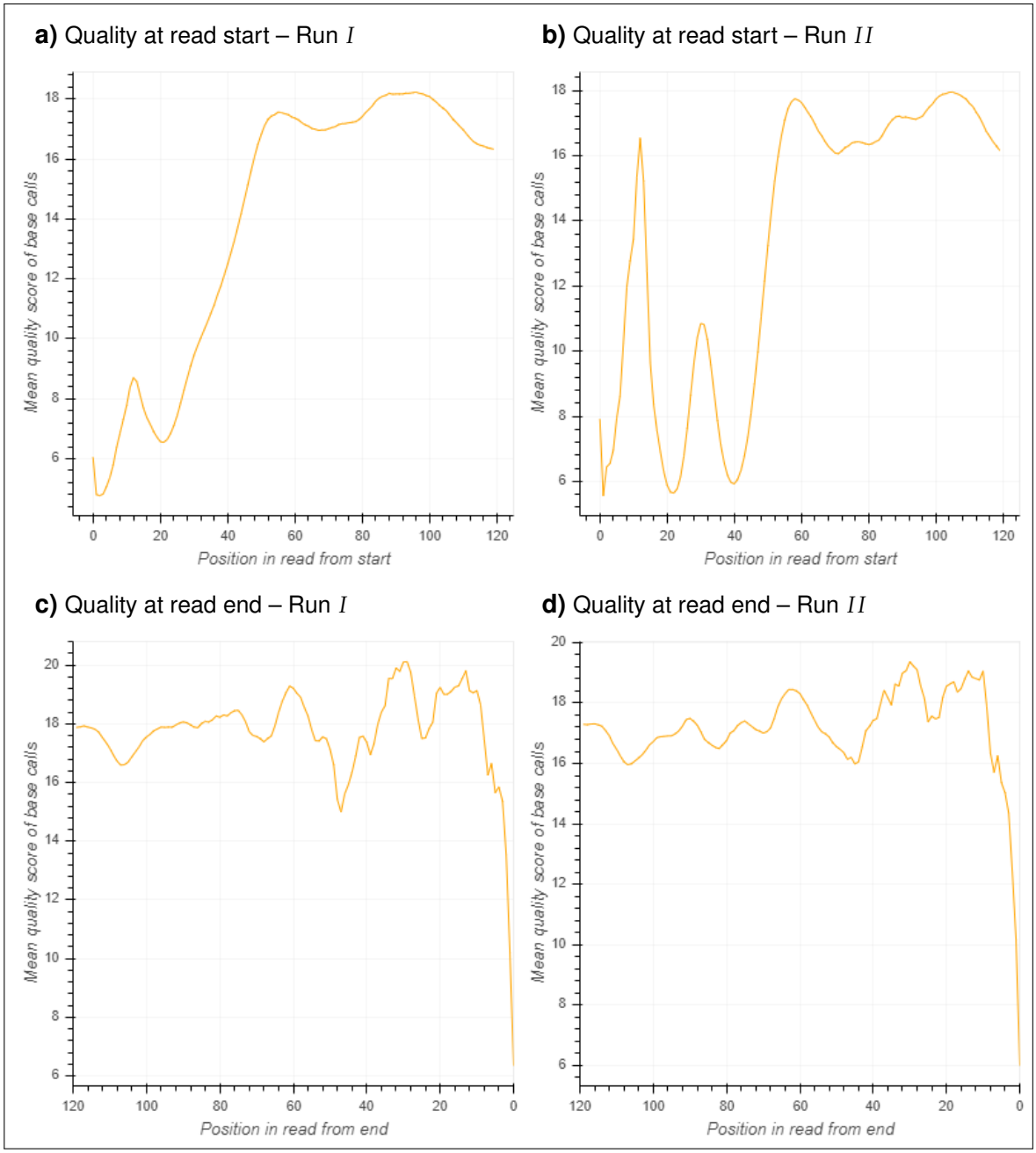

FIGURE S3. Positional quality scores: Representation of the quality scores of the first (Subfigure **a)** and **b)**) and last (Subfigure **c)** and **d)**) 120 bp of all basecalled reads for run I (Subfigure **a)** and **c)**) and run II (Subfigure **b)** and **d)**).

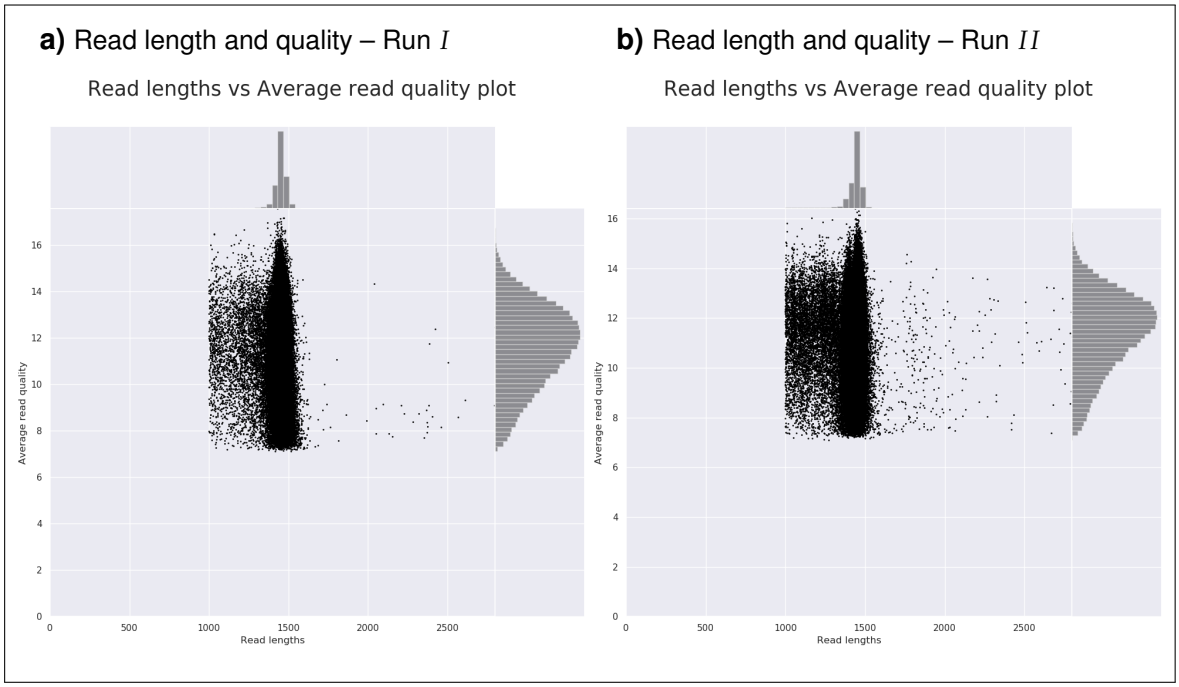

FIGURE S4. Read length and quality distribution after filtering: A scatter plot of the average read quality (given as Phred score) over the read length for run I (Subfigure **a)**) and run II (Subfigure **b)**). The panel above the scatter plot shows the read length distribution histogram, the panel on the right a histogram of the quality distribution. The read length was cut off at 2800 bp.

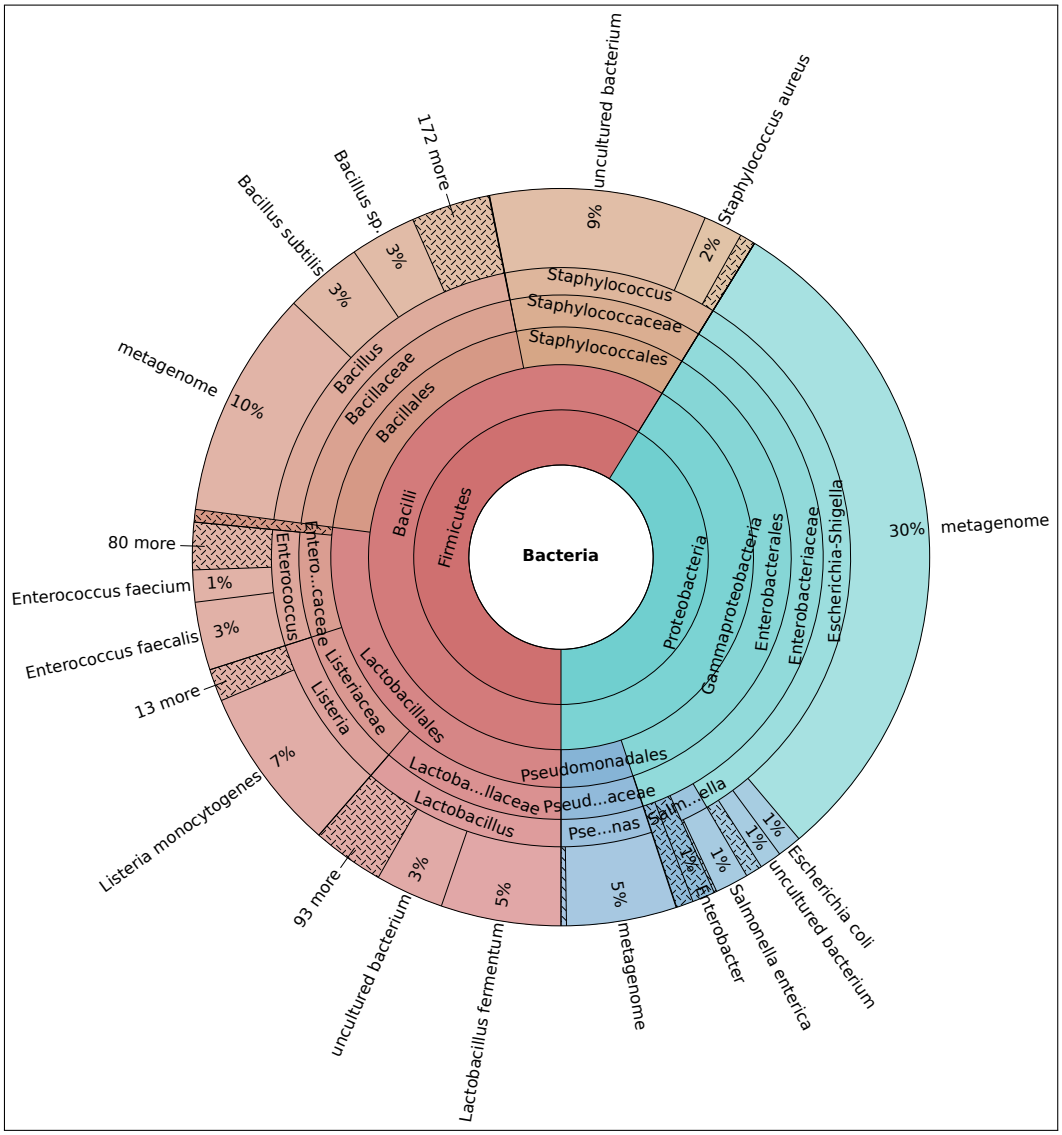

FIGURE S5. Mock community taxonomic composition, ZyCell, run I – k-mer mapping: Multi-layer pie chart of the taxonomic composition for the mock community, shown for ZyCell of run I using k-mer mapping and reestimation of abundance. See Figure 9 for details.

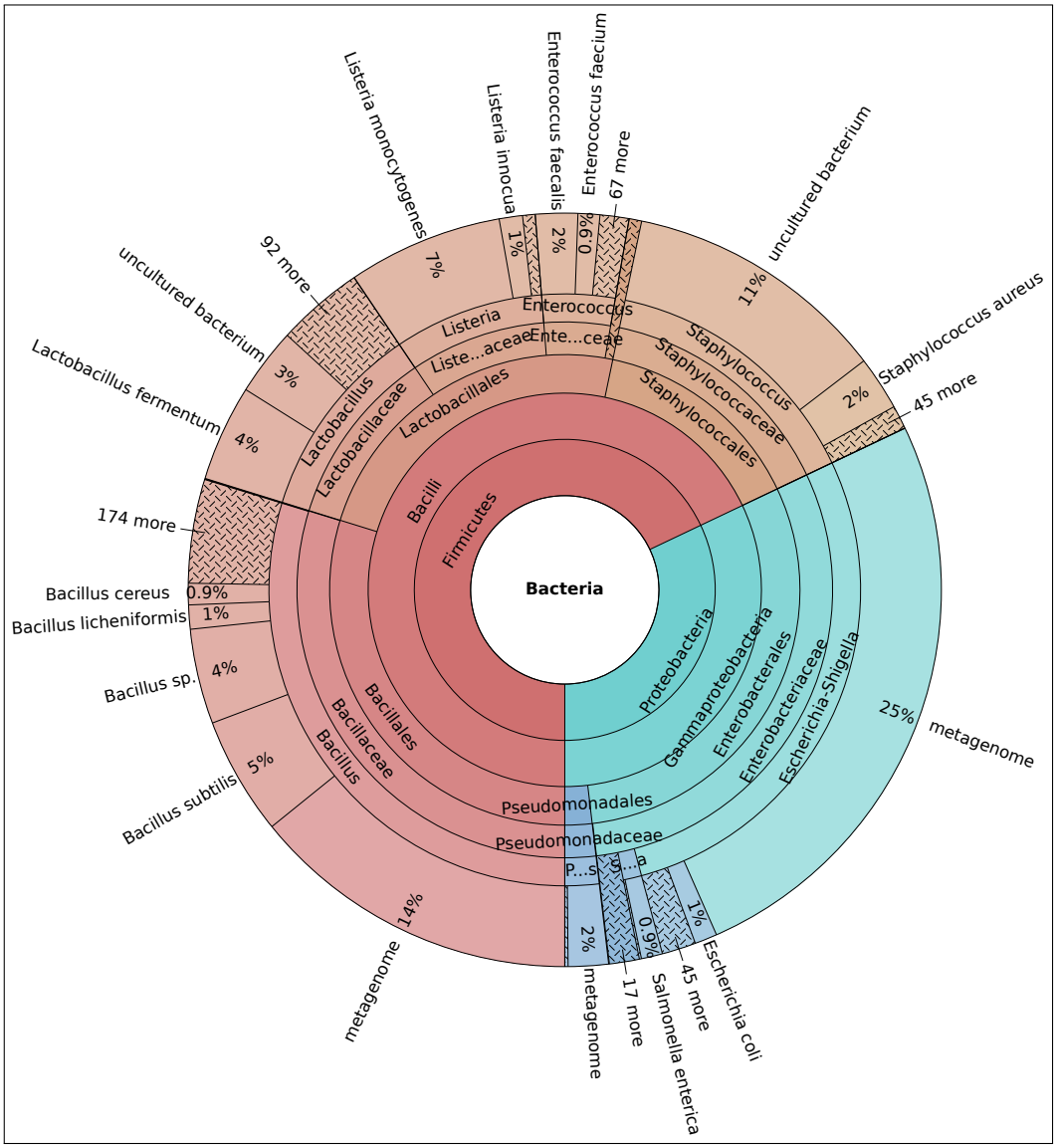

FIGURE S6. Mock community taxonomic composition, ZyDNA, run II – k-mer mapping: Multi-layer pie chart of the taxonomic composition for the mock community, shown for ZyDNA of run II using k-mer mapping and reestimation of abundance. See Figure 9 for details.

FIGURE S7. Mock community taxonomic composition, ZyCell, run I – alignment: Multi-layer pie chart of the taxonomic composition for the mock community, shown for ZyCell of run I using full-length alignment. See Figure 12 for details.

FIGURE S8. Mock community taxonomic composition, ZyDNA, run II – alignment: Multi-layer pie chart of the taxonomic composition for the mock community, shown for ZyCell of run I using full-length alignment. See Figure 12 for details.

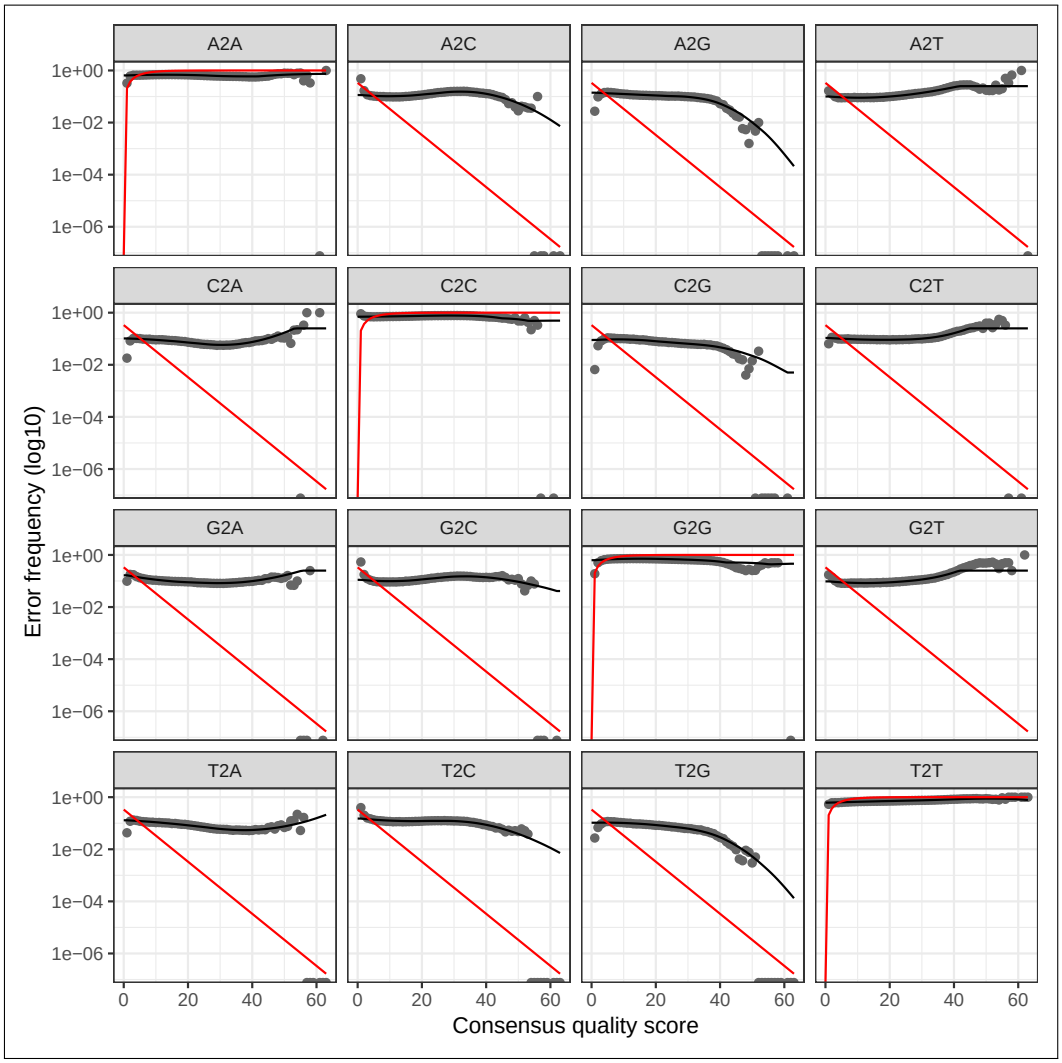

FIGURE S9. Error frequencies in ASV error model: Shown are the error frequencies over the consensus quality score (dots), the expected dependency (red lines) and the model dependency (black lines).

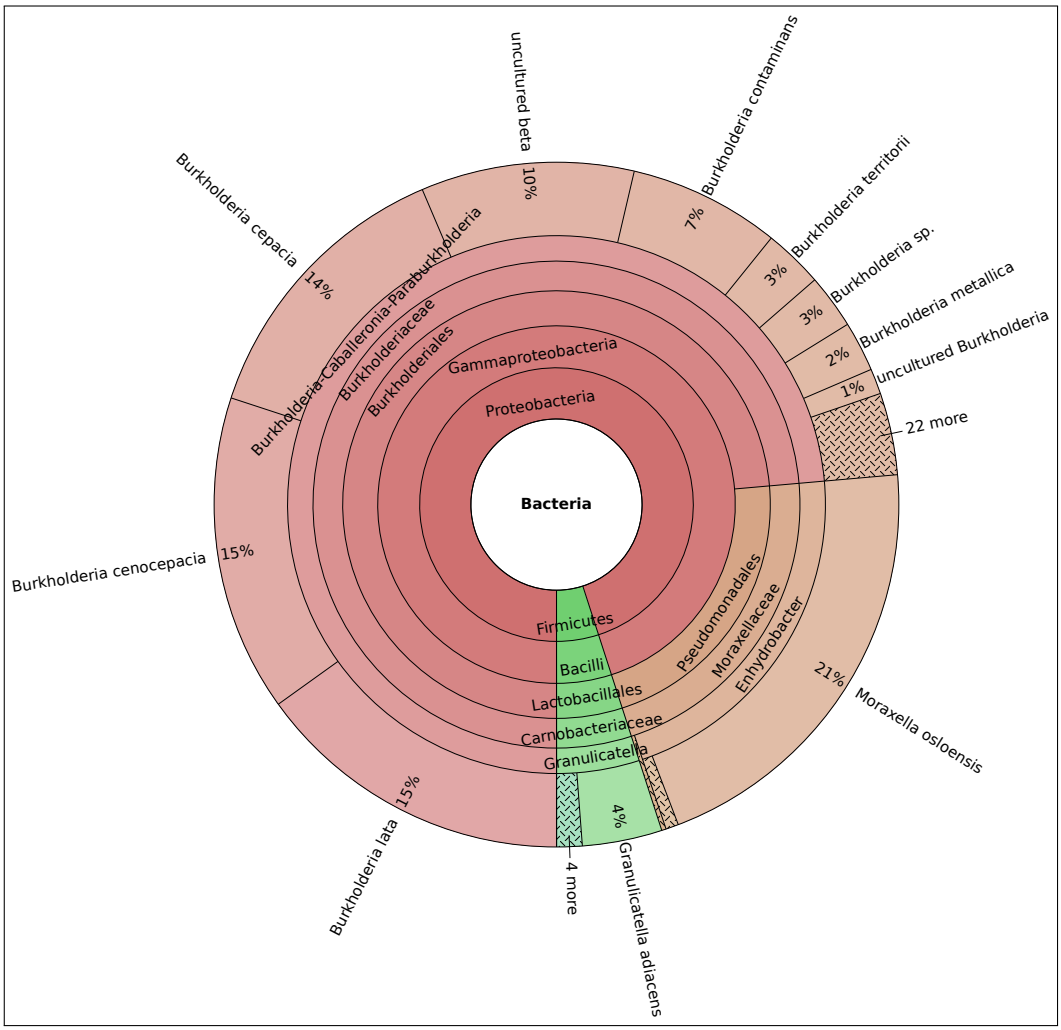

FIGURE S10. No-template control taxonomic composition, run II : Multi-layer pie chart of the taxonomic composition for the no-template control of run II using full-length alignment. See Figure 12 for details.

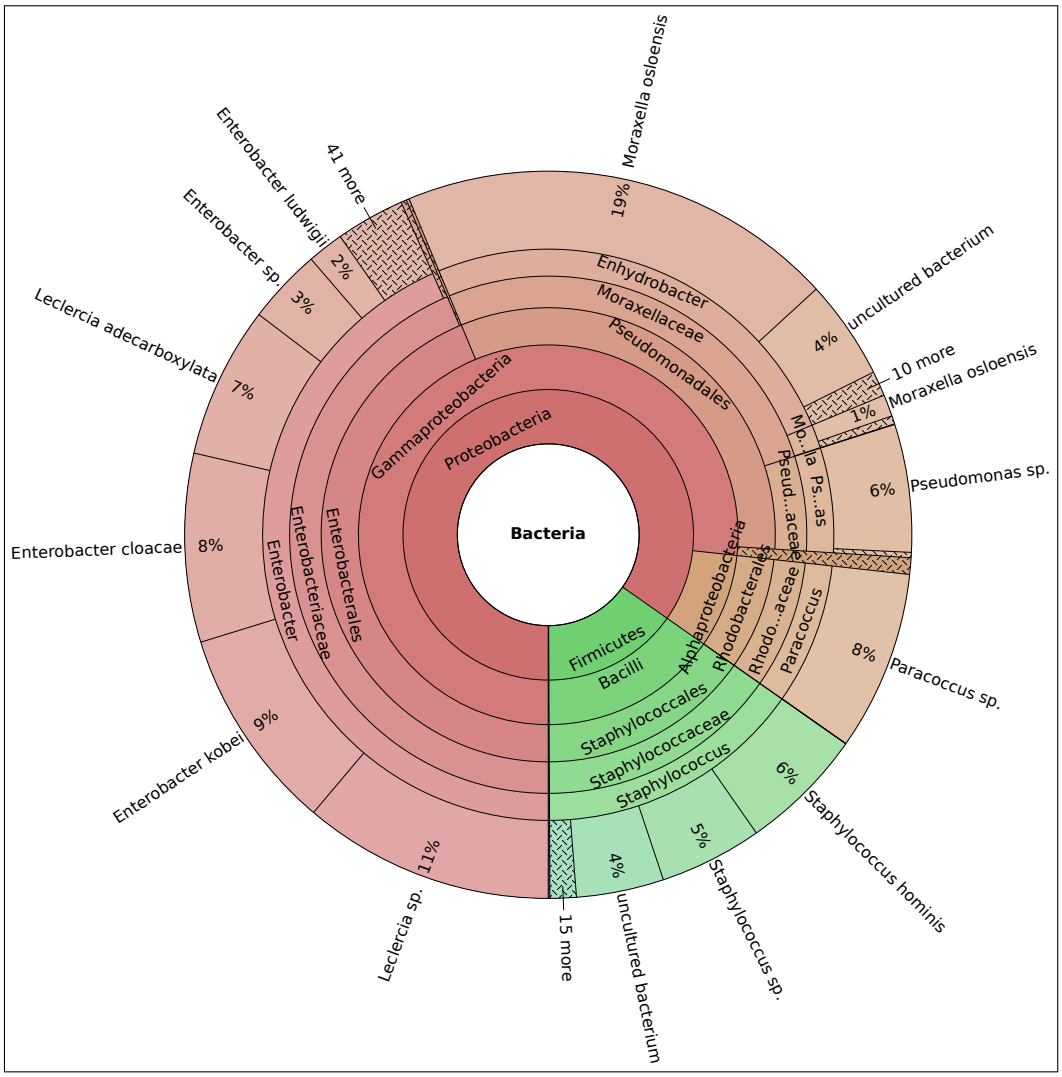

FIGURE S11. Extraction control taxonomic composition, run II : Multi-layer pie chart of the taxonomic composition for the extraction control of run II using full-length alignment. See Figure 12 for details.

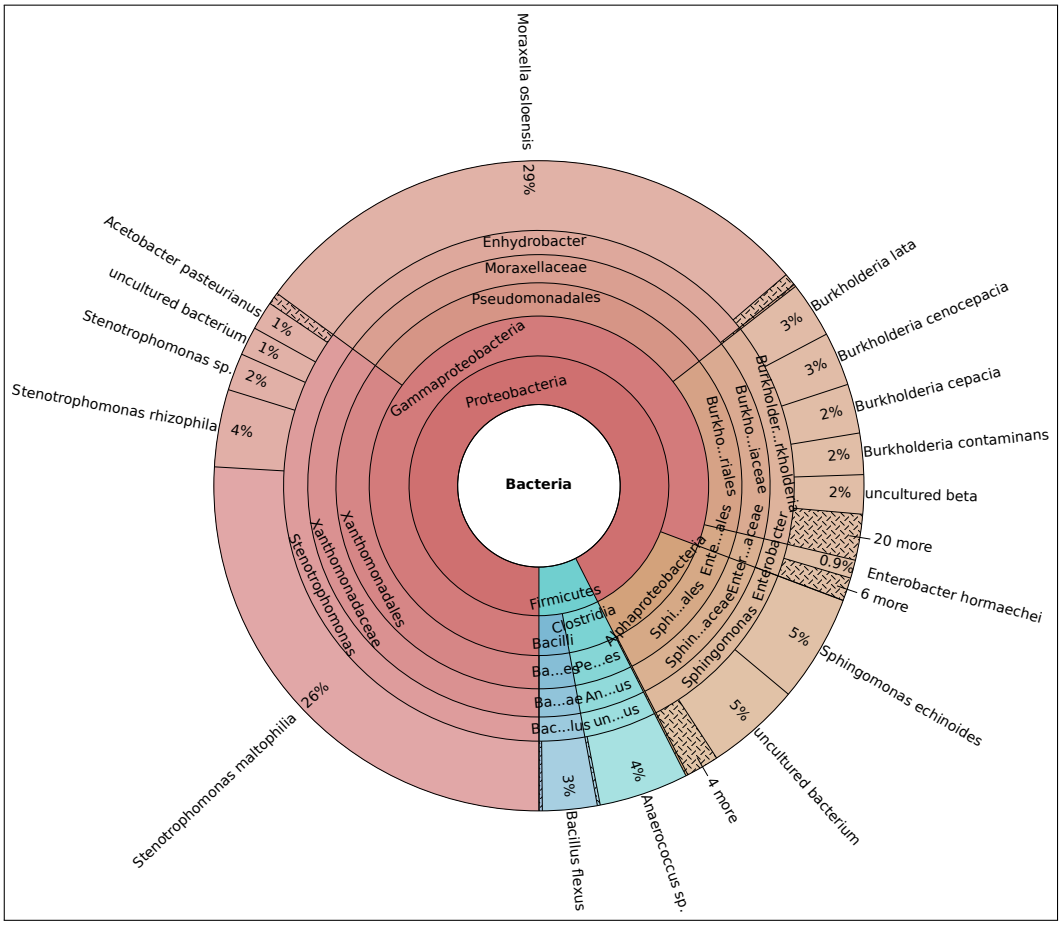

FIGURE S12. Tissue microbiome taxonomic composition – patient Lu05 enritchment batch I**:** Multi-layer pie chart of the taxonomic composition for the lung microbiome, exemplarily shown for patient Lu05 using full-length alignment. See Figure 19 for details.

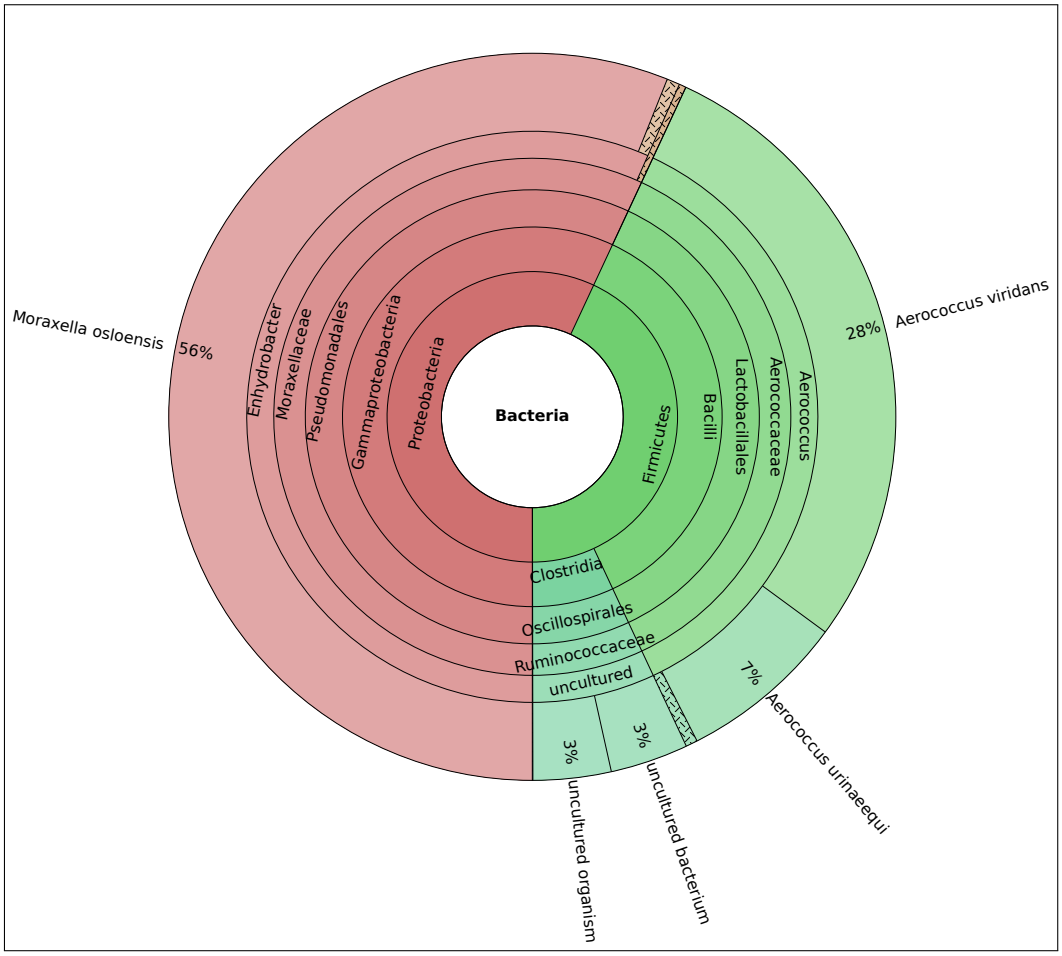

FIGURE S13. Tissue microbiome taxonomic composition – patient Lu05 enritchment batch II**:** Multi-layer pie chart of the taxonomic composition for the lung microbiome, exemplarily shown for patient Lu05 using full-length alignment. See Figure 19 for details.