Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Amplicon clustering methods and the detection of core microbiota in honey bee gut

Annbjørg Helene Nygaard Barbakken
Master in Bioinformatics

Acknowledgements

I want to express gratitude towards my supervisor Lars-Gustav Snipen, for his advice and knowledge during the work with this thesis. I would also like to thank Knut Rudi and Jane Ludvigsen for providing the honey bee data set. The work with this thesis was carried out between fall 2018 and summer 2019, as part of my master's degree in Bioinformatics at the Norwegian University of Life Sciences.

Lastly, I want to thank my partner, friends and family for their guidance and support.

Ås, July 2019

Annbjørg H. N. Barbakken

Abstract

Recent years decline in honey bee populations has lead to an increased interest in the study of their microbiota. In humans, a disturbance in the healthy gut microbiota is linked to several diseases, and because the host-adapted microbiota in the honey bee gut resembles that of mammals, it is assumed that bee gut microbiota also affects the health of bees. This leads us to the study of the core microbiota present in honey bee gut. Gaining knowledge about the core microbiota can help us understand what makes a healthy honey bee.

One route to acquire knowledge about the microbiota is by amplicon sequencing of the 16S rRNA gene. After sequencing, it is common to apply a clustering step to the data. Clustering methods can have a high impact on the results; the main focus in this thesis has therefore been to look at the effects of clustering methods in the study of amplicon reads. Three clustering methods and a control method were used to group amplicon sequences to compare the differences of the clustering results and their ability to detect core microbes.

The results show considerable differences between methods, both in cluster composition and in the detection of core microbiota. One of the clustering methods were not able to detect any core clusters (i.e., clusters part of every sample), and overlooked unique sequences present in a large number of samples.

Two methods did detect core microbiota, consistent with the core genera detected in previous studies. Besides, there were only detected minor differences in the core microbiota composition between different sampling factors such as time of year or gut part.

Results from this study clearly illustrate the importance of method when clustering amplicon reads. Depending on the choice of method, a study could end up with opposite conclusions regarding core microbiota.

Sammendrag

Honningbiepopulasjonen har vært nedadgående de siste årene, dette har ført til økt interesse rundt det å studere mikrobiotaen deres. Hos mennesker er forstyrrelser i den friske tarm-mikrobiotaen koblet til utviklingen av flere sykdommer. Og siden den vertstilpassede mikrobiotaen i tarmen hos bier viser likhetstrekk med den tilhørende pattedyr, kan det antas at tarm-mikrobiotaen hos bier også påvirker helsen deres. Dette gjør det interessant å studere kjerne-mikrobiotaen i tarmen hos bier. Økt kunnskap rundt kjerne-mikrobiotaen kan hjelpe oss å forstå hva som gjør friske bier friske.

Kunnskap om mikrobiota kan tilegnes på flere måter, og en av dem er ved amplicon-sekvensering av 16S rRNA genet. Det er vanlig å anvende clustringsmetoder på slike data etter sekvensering. Slike clustringsmetoder kan ha stor effekt på resultatene, og hovedfokuset i denne oppgaven har derfor vært å se på effekten av clustringsmetode i amplicon-sekvensstudier. For å gruppere amplicon-sekvenser ble det brukt tre clustringsmetoder og en kontrollmetode, og resultatene fra disse ble brukt til å sammenligne effekt av metode på sammensetning av cluster og metodenes evne til å finne kjerne-mikrobiota.

Resultatene viser betraktelige forskjeller mellom metodene, både når det gjelder sammensetning av cluster og hvordan de finner kjerne-mikrobiota. En av clustringsmetodene viste svært dårlig evne til å gjenkjenne kjernecluster (cluster som er i alle prøver), og overså også unike sekvenser som var tilstede i en stor andel av prøvene.

To av metodene fant kjerne-mikrobiota som stemte overens med funn gjort i tidligere studier. I tillegg til dette ble det undersøkt om det var noen effekt av faktorer i forbindelse med prøvetakingen, for eksempel tid på året eller tarmdel. Resultatene viste bare små forskjeller mellom disse faktorene.

Denne oppgaven illustrerer viktigheten av metode ved clustring av amplicon-sekvenser. En studie kan potensielt ende opp med motsigende konklusjoner vedrørende kjerne-mikrobiota, avhengig av hvilken metode som er valgt.

Contents

Chapter 1

Introduction

1.1 Microbial communities

Microbial communities can in general be defined as the organisms interacting with each other and living together in a contiguous environment (Konopka, 2009). When talking about microbial communities, there are two main terms that is frequently used; microbiome and microbiota. The microbiome is the set of genomes contained in the microorganisms in an environment (Boon et al., 2014). The microbiota is defined as the microorganisms, including bacteria, archaea, eukaryotes and viruses, that reside in a specified environment (Sommer and Bäckhed, 2013).

In humans, the intestinal bacteria are thought to be essential for several aspects of host biology such as the metabolism of indigestible polysaccharides, differentiation of the intestinal epithelium and immune system, and the bacteria protect against invasion by opportunistic pathogens (Sommer and Bäckhed, 2013). The microbiota in an environment can be very beneficial to the host or the given surroundings. But microorganisms can also cause disease, either in that an organism itself is causing sickness or that its absence cause the host to develop decease (Khosravi and Mazmanian, 2013). Because microbial communities are ubiquitous and have great impact on their surrounding environments, it is important to gain insight and knowledge about how they work and influence the world around them.

Traditionally, cultivation was the only way to study the composition of microorganisms in microbial communities. However, this approach is very limited because most microorganisms cannot grow in a laboratory setting (Schloss and Handelsman, 2005). Most of the early knowledge about bacterial physiology was limited to the ones that could be grown in a nutrient rich medium. The development of high-throughput sequencing has led to the establishment of the field of metagenomics, which is defined as directly analyzing the genetic contents of an environmental sample without the need for prior cultivation (Oulas et al., 2015). High-throughput sequencing makes the study of whole microbial communities possible in a faster and more cost-effective way (van Dijk et al., 2014). With the ability to sequence large amounts of DNA in an environmental sample, it is possible to access organisms that before was inaccessible (Schloss and Handelsman, 2005).

Sequencing technology first started in the late 1970s, when Sanger et al. (1977) introduced a method of sequencing by chain termination. Sanger sequencing was the most prevalent sequencing technology for 30 years, and the technology resulted in the first complete human genome sequence in 2004 (van Dijk et al., 2014; Collins

et al., 2004). After this it was clear that technologies that gave higher throughput, were faster and cheaper needed to be developed for these types of studies. This is were the next generation sequencing technologies (NGS) come in (van Dijk et al., 2014). Several technologies were developed, with the first being the 454 pyrosequencing method in 2004 (Margulies et al., 2005), succeeded by Illumina, SOLiD, and Ion Torrent PGM technologies in the followed years (van Dijk et al., 2014). These methods are called second generation sequencing technologies today. In the beginning all these methods produced relatively short reads, but development in machinery, base-calling software, and sequencing chemistry have led to the production of longer reads (van Dijk et al., 2014). This development made *de novo* genome assembly and metagenomics possible with Illumina, going from 35 bp long reads to several hundred bp long reads.

Today, the Illumina sequencing platform is the most commonly used sequencing technology. It is currently the sequencing technology with the highest high-throughput per run and the lowest cost per sequenced base (van Dijk et al., 2014). With Illumina, systematic errors are generally low because of the greater coverage/yield (Oulas et al., 2015). The most common errors introduced are substitution errors, and reads where the ends have lower read quality (Schirmer et al., 2015).

Other sequencing technologies, called the third generation, are also being developed (e.g. PacBio and Oxford Nanopore). These can produce substantially longer reads than the second generation technologies, but are still generating a larger error rate than e.g. Illumina, and are still expensive (van Dijk et al., 2014).

1.2 Taxonomic profiling

One goal of a metagenomic analysis is to get a taxonomic profile of the microbial community. A taxonomic profile gives information about what the community contains, e.g. what bacterial taxa the community consists of.

There are two main metagenomic methods used to obtain taxonomical information about a community; marker gene amplification and whole genome shotgun sequencing (Hugerth and Andersson, 2017; Quince et al., 2017).

Marker Gene Amplification

In marker gene analyses, a marker gene is used to identify the different taxa in a sample. This type of community analysis is often used to attain a high-level, but low resolution overview (Knight et al., 2018). The most common marker gene for obtaining information about bacteria in a sample is the 16S ribosomal RNA (rRNA) gene, which codes for the small subunit rRNA of the bacterial ribosome. It was first discovered in 1977 by Woese and Fox (1977) that this gene could infer phylogenetic relationships between prokaryots. The 16S rRNA gene is $\sim$1500 bp long and contains nine variable regions, named V1-V9, with more conserved regions in between (Hugerth and Andersson, 2017; Van de Peer et al., 1996). The variable regions can be used for the discrimination of bacterial species. The short read technology that is frequently used today is not able to sequence the whole 16S rRNA gene, therefore only some of the variable parts of the gene is used in amplicon sequencing. Different variable regions distinguish between different bacterial taxa,

and not all regions are beneficial to use in all types of studies (Chakravorty et al., 2007).

Today, the 16S marker continues to be a frequently used gene to identify bacteria, because all prokaryote species have it in their genomes and also because of all the knowledge and reference sequences collected over the years. This accumulated knowledge makes it impractical to switch to another marker gene (Hugerth and Andersson, 2017). It is important for the taxonomical identification of complex samples that a database contains a large amount of sequences with correct classification and high quality sequences, to determine phylogeny as correct as possible (Hugerth and Andersson, 2017). Marker gene analysis is a convenient way of performing taxonomic classification in large and complex samples, and with Illumina technology it is also inexpensive (Escobar-Zepeda et al., 2015). Some disadvantages with amplicon sequencing is that the 16S rRNA gene generally gives low resolution on a species level (Weinstock, 2012), and many species can have more than one copy of the gene in their chromosomes (Větrovský and Baldrian, 2013) which can lead to different abundances.

Before sequencing, the 16S rRNA gene needs to be amplified using polymerase chain reaction (PCR). PCR depends on primers of short DNA molecules (usually 15-30 bp) to attach to the sequences that are being amplified, for polymerase to elongate the sequence. The conserved regions in the 16S rRNA gene can be utilized for primer design. This need for primers can lead to primer bias in the amplification. The bias is introduced if the marker lacks complementarity to the primer and therefore is not amplified. Primer bias can lead to organisms ending up as false negatives in the microbiome profile (Hugerth and Andersson, 2017). Primer selection can therefore have large impact on the result, and should be selected based on the community being analyzed. PCR can also introduce other artifacts, such as chimeras.
For the amplification of the data in this thesis, the primers used are the 16S rRNA primers designed by Yu et al. (2005). The forward primer is PRK341F (CCTACGGGRBGCASCAG) and the reverse is PRK806R (GGACTACYVGGGTATCTAAT).

The next step is sequencing, and for the data in this thesis, the Illumina sequencing platform was used. The reads that are output from the sequencing are called amplicons, which refers to the PCR amplification step. The Illumina technology produce paired-end reads as output, this makes it possible to obtain longer reads and also have more certainty of the sequence in the regions where the reads overlap. Based on the errors produced by both PCR and Illumina there are different filtering steps performed on the amplicons. After filtering the amplicons are clustered into operational taxonomic units (OTUs), based on similarity, in an attempt to eliminate remaining erroneous sequences (Hugerth and Andersson, 2017). These erroneous amplicons formed by PCR and sequencing are expected to deviate from the true sequence only by a few bases. Several algorithms are developed to perform this step, containing different approaches and results. In this thesis the focus is on the identity clustering algorithm implemented in the VSEARCH software (Rognes et al., 2016), the Dada2 algorithm (Callahan et al., 2016), and the Swarm v2 algorithm (Mahé et al., 2015). The VSEARCH approach expects a predetermined identity threshold. This threshold is most often set to 97% sequence identity as species cut-off, but this is argued to be arbitrary as a species limit (Koeppel and

Wu, 2013). Callahan et al. (2016) argues that Dada2 produce amplicon sequence variants (ASVs), which is thought to be more exact than OTUs, but for simplicity all clusters are referred to as OTUs.

To be able to compare communities, it is common to assign taxonomy to the OTUs. There are multiple algorithms created to assign taxonomy, and all attempts to classify as fast and correctly as possible. Several databases (e.g. SILVA, RDP) are also constantly updated because new sequences are discovered frequently, and as mentioned earlier a good classification depends on a good database (Hugerth and Andersson, 2017). The taxonomical classification results in a taxonomic profile for each sample, often based on genus level.

Whole Genome Shotgun Sequencing

The other common method to attain a taxonomic profile for a set of samples is by whole genome shotgun (WGS) sequencing. This is an untargeted method, sequencing all genomes of microbial organisms present in a sample (Quince et al., 2017). There are no amplification of a specific marker gene, as all the DNA in a sample is sequenced directly with e.g. the Illumina sequencing platform. Thus, there will not be any biases introduced by PCR in the data. The WGS approach is more expensive than amplicon sequencing in preparation, sequencing and analyses of samples, but on the other hand it yields more detailed genomic information and taxonomic resolution (Knight et al., 2018). And with WGS sequencing there is no need for any previous knowledge about the community, in contrast to amplicon sequencing where some assumptions have to be made in the selection of primers.

The easiest way to identify which bacteria a WGS sample contains is to directly compare it to a database with already characterized bacterial sequences. As mentioned for the amplicon data, a well characterized and maintained database is crucial for the quality in this type of data as well. The main limitation for this method is that previously uncharacterized microbial organisms are difficult to profile (Quince et al., 2017). For environments that consists of microbes well represented in databases of reference sequences, however, the taxonomic profiling can differentiate bacteria with species level resolution. One such example is the human gut microbiome, where it over the last years have been identified a large amount of sequences (Zou et al., 2019). The run time can be quite large with the number of reads produced by WGS sequencing, and heuristic search algorithms are developed to speed up the taxonomical classification.

With WGS data it is also possible to assemble the shorter reads into longer contigs of DNA in an attempt to assemble the full genomes of the organisms in a sample. This approach is more computationally expensive than the assembly-free approach, however the assembled genomes can lead to new insights and reference genomes (Knight et al., 2018). A taxonomical profile can then be obtained from the contigs. The assembly can be more difficult if the microbial community is highly complex with high microbial diversity, this can yield fragmented assemblies and disturb taxonomical classification (Knight et al., 2018). Soil samples are an example of a community that is hard to assemble because of high diversity and low sequencing coverage (Howe et al., 2014).

1.3 The core microbiota

Because prokaryotes are so important to us, it is of interest to understand community characteristics regarding microbial composition. Core microbiota is a term often used, and it refers to the core microbes in a specified environment. These core microbes are always present and are possibly of great importance. An example of this is the human gut and finding the microbiota common in all human individuals (Turnbaugh et al., 2007). If we get more knowledge about what a "normal" microbiota looks like, e.g. the healthy human gut microbiota, we can then use this information to treat the microbiota in dysbiosis, microbial imbalance. Imbalance in the microbiota can make the host more susceptible to infection, it can lead to obesity, it is linked to Crohn's disease and the development of type 2 Diabetes (Khosravi and Mazmanian, 2013; Manichanh et al., 2006; Larsen et al., 2010).

To characterize the core microbiota, bacterial sequences need to be sequenced. Given that the technology we have today does not sequence a sample perfectly, it will always introduce bias into the data. Low sequencing depth and primer bias can also lead to prokaryotes being overlooked in the sequencing, hence they are not part of the later analyses. A critical point in core studies is the determination of the threshold deciding what prevalence a microorganism must have to be a part of the core microbiota. A bacteria present in 100 % of the samples, but only present in 90 % of the sequenced data will not be part of the final core microbiota if the threshold is 100 % prevalence. It is important to be mindful of this when the prevalence threshold is set. If the threshold value is too low there could be false positives, but if the value is too high there could be false negatives.

Another aspect that can influence the detection of core microbiota is the clustering level of the 16S rRNA gene amplicons. The number of OTUs will vary depending on the level, with a more stringent clustering level there will be more OTUs than with a less stringent clustering. When there are a high number of OTUs the amplicons get divided into more groups, hence there will be fewer amplicons per OTU. With fewer OTUs the amplicons does not have to be divided as much and there will be more amplicons per OTU. If OTUs that are actually different species are lumped together because of a low clustering threshold, there is a higher possibility that this cluster will also have more reads across samples and therefore be detected as part of the core. In an article by Aguirre de Cárcer (2018) they show that different numbers of core bacteria are obtained with varying clustering thresholds. That study used 100 % prevalence as core threshold, and obtained a higher number of core OTUs when the clustering threshold where less stringent.

1.4 The honey bee gut

The data used in this thesis originates from the gut of honey bees. The host-adapted gut microbiota in honey bees resembles the gut microbiota in mammals, but the composition is more simple in the bee gut (Kwong and Moran, 2016). Honey bees are important contributors both to human food consumption and to many ecosystems, because of their role in pollination of different fruits, vegetables, and wild flowers (Genersch, 2010). Recent years have shown a decline in honey bee populations all over the world, and there are several factors thought to be drivers of the declining populations (Sánchez-Bayo and Wyckhuys, 2019). One of them being infections

caused by pathogens and parasites. As for humans, the gut microbiota of bees are thought to protect from pathogens by hindering their colonization (Anderson et al., 2011). Thus, the study of the gut microbiota may lead to insight in one of the reasons for the declining populations.

The honey bee gut is divided into four parts; the crop, the midgut, the ileum, and the rectum, which contains differing abundances of bacteria (Kwong and Moran, 2016). The crop contains very few bacteria, and is used to store and transport nectar for the purpose of feeding larvae and the production of honey (Martinson et al., 2012). Few bacteria is found in the midgut as well. The inside lining is continuously shed when meals pass, thus inhibiting the attachment of bacteria (Martinson et al., 2012). Food are digested and absorbed in the midgut of the honey bees. A higher abundance of bacteria is found in the ileum, where microbial attachment is much easier (Martinson et al., 2012). This gut part is much smaller than the midgut, but still contains more microbes. Those nutrients that were not absorbed in the midgut can be collected in the ileum. The rectum also contains a high abundance of microbiota, as it is a stable nutrient-rich environment (Martinson et al., 2012). The contents in the rectum are digested waste waiting to be disposed of.

When the bees emerge from the pupal stage, their guts consist of few to no bacteria (Powell et al., 2014). Contact with other bees and the hive ensures development of a normal core gut microbiota. This can lead to assumptions that different hives have slightly diverse bacterial strains, because the bees only acquire their core microbiota from their own hive environment.

1.5 Aim of the study

In this thesis the focus is on data collected from the microbial community of the honey bee gut. A total of 460 samples have been collected over 9 months, from three different hives. For all samples, a portion of the 16S rRNA gene have been amplified and sequenced. From this, more knowledge about the core microbiota of honey bees could be obtained.

The primary aim of this study is to see how various bioinformatic methods affect the estimation of the core microbiota in general. It is of interest to see how different clustering methods and thresholds influence the detection of core microbiota. If the method used makes a big difference in what bacteria is defined as the core, it is important to learn how these methods affect the result and how to best avoid a wrong determination of core taxa. Based on these results, the secondary aim is to uncover new insight to the core microbiota of honey bees. Recent years have shown a decline in honey bee populations and insight in the gut microbiota can lead to solutions that prevent further decline.

Chapter 2

Methods

2.1 Data

The data set used in this thesis is a set of bacterial 16S rRNA gene amplicons sampled from the gut of honey bees. The bees were sampled from three different hives (K2, K3, K5) over the course of 9 months, and 10 samples were taken from each hive each month. In March and April the bees were sampled from hive K2 and K3, and in June, July, August, and November the bees were sampled from hive K3 and K5. The hive change from K2 to K5 was a consequence of the queen leaving the hive (K2) in May. The gut of the bees are divided into four different parts called Crop, Midgut, Ileum and Rectum. This creates a total of 460 samples (table 2.1). There were no samples from the Crop in March.

The 16S rRNA genes from the samples where first amplified using targeted polymerase chain reaction (PCR) and then sequenced with Illumina. The forward and reverse primers used are PRK341F and PRK806R designed by Yu et al. (2005), targeting the 16S rRNA gene in prokaryotes.

Table 2.1: A table with the number of samples per category Month, Gut part and Hive. All samples come from one of each of the categories and are a combination of a month, a gut part and a hive.

Month	March	60
	April	80
	June	80
	July	80
	August	80
	November	80
Gut part	Crop	100
	Ileum	120
	Midgut	120
	Rectum	120
Hive	K2	70
	K3	230
	K5	160

Most of the analyses performed in this thesis is performed using the programming language R. R is freely available, and specializes in statistical computing and graphics. R is available for download through: https://www.r-project.org/

2.2 Sequence clustering

The 16S amplicon reads can, based on assumptions of similarity, be grouped in clusters. These clusters can be used to determine which taxa are present in a sample, utilizing taxonomic classification. Advantages of using the 16S rRNA gene for identification of bacterial communities is that it is present across bacterial species and a relatively conserved gene, it contains conserved regions ideal for PCR primers and hyper-variable regions suited for identification of the different bacterial taxa (Hugerth and Andersson, 2017). The amplicon sequences originating from the same genome is assumed to be nearly identical (Větrovský and Baldrian, 2013) and can, therefore, be clustered together. However, sequencing can introduce errors in the amplicon reads, and sequences that are slightly different can still come from the same genome(Hugerth and Andersson, 2017). It is also possible to cluster the sequences with a less stringent approach, and assume that different strains have slightly different 16S genes but still contain in the same taxon and therefore should be clustered together(Hugerth and Andersson, 2017). Different levels of clustering are applied differently depending on the wanted outcome. If the goal of the clustering is to separate based on the strain level, the grouping has to be more stringent than if the separation is based on a higher taxonomic level. A cluster of sequences created by a clustering algorithm are often referred to as an operational taxonomic unit (OTU).

In a study of the core microbiota, the clustering level is crucial. A natural place to start is then to look at the effect of different clustering thresholds on the grouping of sequences. Four main methods were used to cluster the reads in this thesis:

- 97% identity clustering from VSEARCH
- Denoising algorithm from Dada2
- Single linkage clustering with Swarm
- Dereplication (identity clustering with 100% threshold).

For the rest of the thesis, these are referred to as, respectively, Vsearch, Dada2, Swarm, and Dereplication. The Vsearch method was chosen because it represents the most used way of clustering 16S sequences, the Dada2 and Swarm methods were chosen because they offer a completely different algorithm for clustering the reads and they also claim to create a more refined output of OTUs (Callahan et al., 2016; Mahé et al., 2015). The Dereplication method was chosen to see how clustering only the unique sequences would compare to the methods where the sequences are clustered based on a threshold of some sort. The methods and how the clustering is performed is described below for each method separately. Dereplication is a step in the other three methods, and because the procedure is the same across methods, the output is not dependent on the software. The VSEARCH (Rognes et al., 2016) software was used to perform the dereplication to create the output data used for the Dereplication method, and are explained under section 2.2.1.

An OTU table was produced for each method, containing the read count for each OTU in each sample. Every cluster (OTU) is represented by a sequence, this is referred to as the centroid sequence or the centroid. The centroid sequence is often the most abundant sequence from the cluster. A simple illustration of the clustering pipelines are displayed in figure 2.1.

QIIME

The QIIME (Caporaso et al., 2010) package was not used to access any of the clustering methods mentioned above. QIIME is a package containing several methods and pipelines for analyzing microbiome data, and is frequently used for this purpose. QIIME, among many other plug-ins, contains VSEARCH and Dada2.

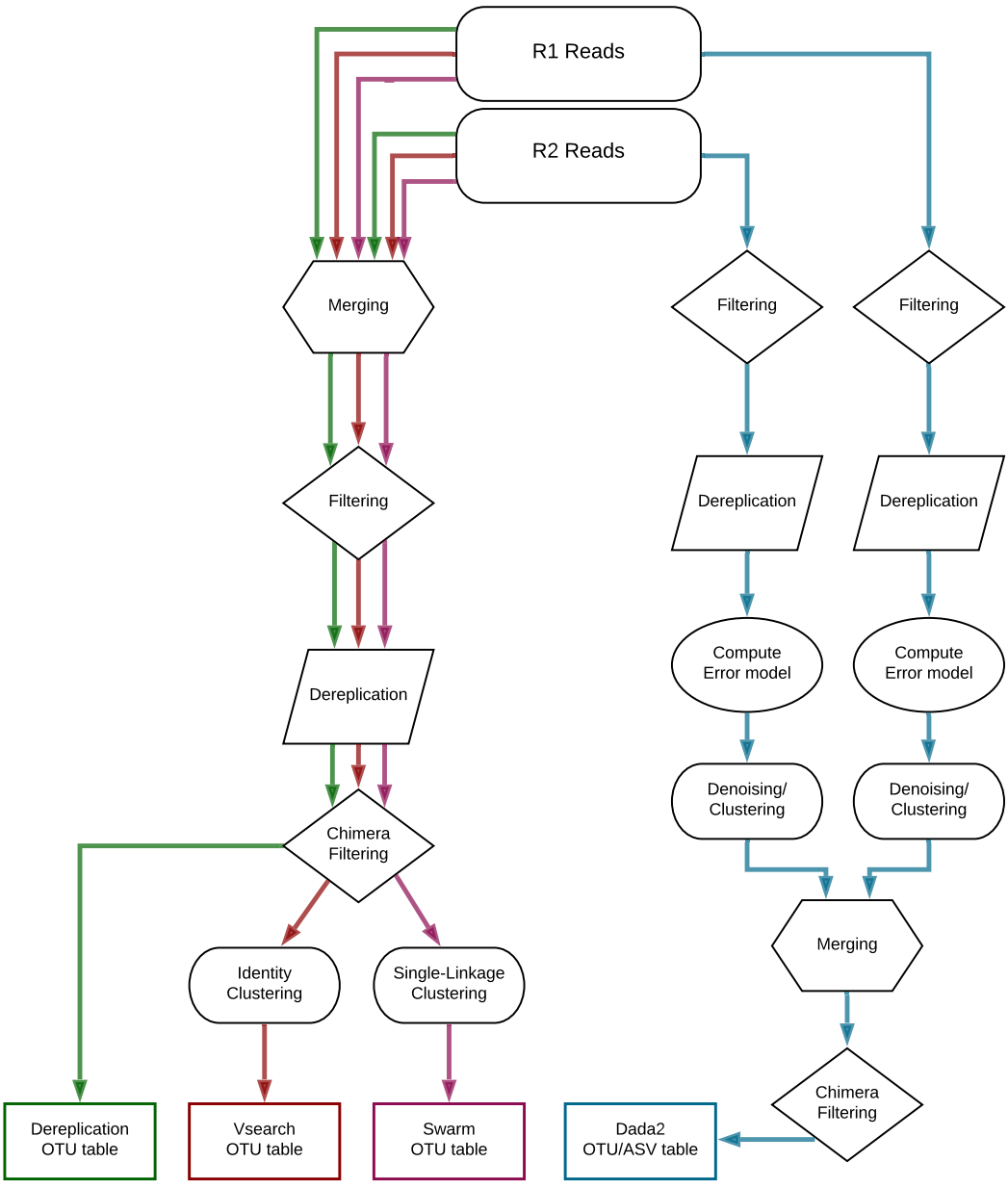


Figure 2.1: Flow chart displaying the general pipeline for the four clustering methods; Vsearch (red), Dada2 (blue), Swarm (pink), and Dereplication (green).

2.2.1 Vsearch

One method to cluster the 16S sequences is by abundance clustering. For this thesis the software VSEARCH (Rognes et al., 2016) was used to perform this type of clustering. The software was also used for merging, filtering, dereplication and chimera detection. As stated in the article by Rognes et al. (2016) the VSEARCH software was developed as an open-source alternative to USEARCH (Edgar, 2010), and it includes most of the functions most commonly used and a few new additions.

Merging

The bacterial reads are paired-end reads and therefore have to be merged to form one sequence. To perform this merging the VSEARCH command *fastq_mergepairs* was used. This uses an algorithm that computes an optimal alignment, without gaps, for the overlapping region between the forward sequence and the reverse-complemented reverse sequence. There are several requirements for the alignment, including a minimum overlap length, a maximum number of mismatches and a minimum and maximum length of the resulting merged sequence. For the merging a match score $(+4)$ and a mismatch score (-5) is used. With the information from the quality score, given by the sequencing machine, the match and mismatch score are weighted by the probability that the sequence match or mismatch, respectively. For the resulting merged sequences VSEARCH computes posterior quality scores for the regions where the sequences overlap. The input for the merging is two Fastq files and the output is one Fastq file with the merged sequences including the quality scores.

Filtering

After merging the sequences goes through a filtering step to discard sequences with errors, low quality scores or sequences that are too short or too long. This step is performed by the VSEARCH command *fastq_filter*. The options specified for the filtering was *fastq_maxee_rate* at 1.0 where sequences are discarded if the expected errors per base is more than the specified number, *fastq_minlen* at 300 where sequences shorter than this length is discarded, *fastq_maxlen* at 600 where sequences longer that this is discarded and *fastqmaxns* at 2 where sequences with more than 2 N's are discarded. The input for this step is the Fastq file from the merging step and the output is one Fasta file containing only the sequences.

It is also of interest to discard PCR artifacts such as the chimera sequences. Chimeric sequences come from the amplification process, and are sequences that stem from two or more of the original sequences in the sample. They are formed when a partially extended sequence from an original sequence is re-annealed to another original sequence in the next cycle of the PCR amplification process. This mechanism is called incomplete template extension and is the most common way that chimeras are produced (Edgar et al., 2011). In VSEARCH there are two ways of detecting the chimeras, either *de novo* with the command *uchime_denovo* or with a reference database with the command *uchime_ref*. The algorithm used by VSEARCH is the same as the one described by Edgar et al. (2011). The general description of how the chimeras are detected with VSEARCH is described by Rognes et al. (2016); each query sequence is divided into four segments and a heuristic search

function is used to look for similarities of each segment with the potential parent sequences. Because the reference based method rely on a proper database, suitable for the given data set, to detect chimeras the *de novo* approach were chosen. With this approach the sequences are only compared to the other sequences in the given data set, and chimeras are decided to be the sequences that originate from two or more of the sequences present in the data set.

Dereplication

After merging and filtering, the clustering process can begin. The first step of partitioning the sequences is dereplication, which is to divide the sequences into clusters of 100 % identical sequences. The VSEARCH command used for this is called *derep_fullength* and this function performs a full-length dereplication using a hash table. At the start of the dereplication the hash table is empty, and for each new input sequence VSEARCH computes the hash and performs a lookup in the hash table. When the lookup results in an identical sequence the input sequence is clustered together with the matching sequence. If there is no identical sequences in the hash table the input sequence is inserted into the table and the dereplication continues.

Clustering

After dereplication is performed, the next step is to cluster the sequences based on abundances and a sequence similarity threshold. The VSEARCH command to perform this type of clustering is called *cluster_size* and the identity (sequence similarity) is adjusted with the *id* option. The VSEARCH algorithm is greedy and heuristic, and performs the clustering *de novo*. *De novo* clustering means that it is performed only using the sequences in the data set when inferring clusters, it is also possible to use a reference database and infer clusters from already known taxa. The clustering of the data set where performed at 97% sequence identity because this is a common similarity threshold for 16S bacterial sequences. When clustering, VSEARCH uses each input sequence as a query and searches against a database of centroids. The centroid database is initially empty and when no matches are found the query sequence becomes a centroid of a new cluster in the database. The first sequence found in the database with a similarity equal to or above the threshold, will be the centroid the query is clustered with. Because the input sequences are sorted by abundance the first centroids in the database will be the clusters containing the most sequences.

2.2.2 Dada2

Dada2 (Callahan et al., 2016) is an R package built for denoising Illumina amplicon reads and it contains functions to perform every processing step from demultiplexed forward and reverse Fastq files to a finished, chimera free sequence variant table. The pipeline is quite different from the Vsearch way of performing the processing and the clustering of the sequences is not based on a percent sequence identity. In the article by Callahan et al. (2016) the authors argue that the algorithm is denoising the reads and producing what they call amplicon sequence variants. These amplicon sequence variants are supposed to represent the true biological sequence of

a species and be a more refined result than the regular identity clustering. Another big difference from Vsearch is that the reads are not merged until after the denoising step. This is to achieve greater accuracy because the denoising algorithm uses the empirical relationship between the quality score and the error rates (Callahan et al., 2016). If the reads are merged prior to denoising the relationship will be different between the forward part, overlapping part and the reverse part of the resulting merged read and this difference will interfere with the algorithm.
Unless otherwise noted the default values for the functions are used. The Dada2 pipeline was run according to the Dada2 tutorial on GitHub (https://benjjneb.github.io/dada2/tutorial.html, read: February 27, 2019) using Dada2 version 1.10.1.

Filtering

The first step in the pipeline is to filter the Fastq files using the *filterAndTrim()* function. This function is based on the *fastq_filter* command from the USEARCH software (Edgar and Flyvbjerg, 2015). The input to the function is a set of corresponding forward and reverse (optional) Fastq files. If the reverse files are provided the filtering is performed independently, however for the read pair to be output, both the forward and the reverse read must pass the filtering. The reads are trimmed to a specified length, 290 and 240 for the forward and reverse reads respectively, and reads shorter than this are discarded. Reads containing Ns are removed, and reads with an expected error above a specified number are also discarded. The maximum expected error threshold was set to 3. The output from the filtering is separate forward and reverse Fastq files containing only the sequences that passed the filtering.

Dereplication

The next step is to perform dereplication using the *derepFastq()* function. This is done on the forward and reverse Fastq files separately. This is different from both Vsearch and Swarm which merges the sequences before dereplication. The input to the function is a Fastq file for each sample and the output is an R object containing the unique sequences and their abundances. In addition the consensus quality scores for each position in the reads are calculated by taking the mean of the quality scores for each unique sequence.

Denoising

The denoising of the sequences by Dada2 depends on an error model, and based on this an error rate for each sequence alignment is calculated. The errors are assumed to occur independently both within a read and between reads. This error model is inferred from the data by the function *learnErrors()* that takes as input the filtered Fastq files and outputs an R object containing the error model. Two error models are inferred from the data, one for the forward reads and one for the reverse reads. The function performing the denoising and partitioning of the reads is called *dada()*. This function takes the dereplicated reads and the error model as input. The unique reads are first grouped together in a single partition and the most abundant sequence is assigned to be the center sequence of the partition. This center sequence is then

compared to all the unique sequences using pairwise sequence alignment which is performed by a vectorized implementation of the Needleman-Wunsch algorithm with ends-free gapping (Callahan et al., 2016). Because this is computationally expensive; heuristics are implemented by default, but these can be altered by the user. When executing the sequence alignments the error rates are calculated and stored. For each unique sequence an abundance p-value are also calculated. The abundance p-value is explained by Rosen et al. (2012) as the probability of having observed at least as many identical reads as observed of each sequence, on the condition that at least one is observed. The abundance p-value will be low if the number of reads of a sequence i are higher than could be explained by the errors introduced during amplification and sequencing of a sample sequence j. Singleton sequences will have an abundance p-value of 1 because it is calculated on the condition that at least one sequence is observed; hence, Dada2 will not allow singletons to form their own partitions. New partitions are formed if the smallest p-value calculated is below OMEGA_A, the default value for this threshold is $1e-40$. The sequence having the smallest p-value will become center of the new partition, the other unique sequences are then compared to this new center sequence. All the unique sequences are placed in the partition most likely to have produced it. The procedure can now iterate and a new unique sequence, with the smallest p-value, can form a new partition, the other sequences are reshuffled and the division continues until there are no sequences with a p-value below the OMEGA_A threshold.

The denoising was performed without pooling the samples, as described in the tutorial on GitHub. Without pooling of the samples the denoising is performed on the samples separately, however the same error models is used for all samples, in the forward and reverse data respectively.

Merging

After denoising the forward and reverse sequences needs to be merged. This is performed by the function *mergePairs()*, which takes as input both the forward and reverse denoised R-objects and the R-objects containing the dereplicated forward and reverse reads. The output is an R-object containing the final merged reads with the representing sequences and abundances. The function will only merge the sequences with exact overlap, because it is expected that the denoising removed most of the substitution errors. The merging is not very well documented, and it is difficult to understand the exact procedure.

When the merging is performed, an OTU table is constructed containing all sequence variants/OTUs and samples with the corresponding read counts.

Chimera filtering

Dada2 also contains a function to remove chimeras, this is called *removeBimeraDenovo()*. For this data set the removal was performed on the sequence table after denoising and merging. The function is meant to be used on denoised reads, and are therefore more sensitive than other methods to remove chimeras (Callahan et al., 2016).

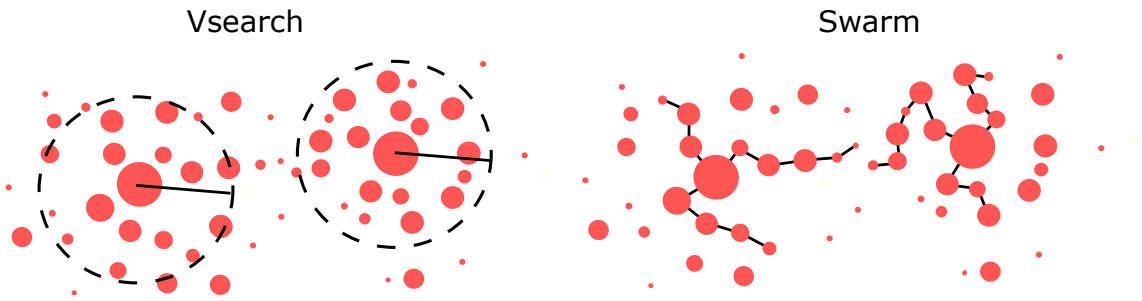


Figure 2.2: Simplified illustration of the clustering algorithms Vsearch and Swarm. Vsearch cluster sequences with a global similarity threshold, Swarm cluster sequences with a local similarity threshold. Each circle represent a unique read, and the size represent the abundance.

2.2.3 Swarm

Swarm (Mahé et al., 2015) is a *de novo* clustering algorithm that use a local threshold, which is different from Vsearch that use a global threshold for clustering. The Swarm software only contains the clustering algorithm; hence, the VSEARCH software was used to filter, merge, and dereplicate the sequences before clustering. Settings and procedures used for the Swarm data are the same as for the Vsearch data, and are explained under section 2.2.1 (Merging, Filtering, Dereplication). Swarm is similar to Dada2 in that it does not require a user-determined, fixed sequence similarity threshold like the Vsearch method with e.g., 97% sequence identity. Another difference to the Vsearch method is that the input order of sequences does not influence the clustering. For the clustering of this data set the Swarm v2 was used, the d value was set to the default value of 1, and the fastidious option was enabled. The d is set to 1 to obtain the finest possible partition of the OTUs in the data set.

Clustering

Swarm clustering is performed in two phases; the growth phase and the breaking phase. The clustering algorithm takes dereplicated Fasta files as input. In the growth phase the first step is to create a pool of amplicons from the input sequences, and make the first available amplicon an OTU seed. Then Swarm generates all possible "microvariants" for this seed and uses a hash table to see if these microvariants are present in the remaining pool of sequences. This is an exact-string comparison approach instead of using a pairwise alignment approach. Mahé et al. (2015) defines a microvariant as, only when d=1, a sequence with one difference to the original sequence. This difference can be caused by either a deletion, an insertion, or a substitution. The sequences in the pool with a match to any of the microvariants are added to the cluster and become subseeds. The subseeds are then compared to the remaining sequences in the pool in the same way. This process of comparing and adding sequences to the cluster is iterated until there is no more sequences added, and the cluster becomes closed. The process starts over with selecting a new amplicon as a seed sequence for an OTU, adding subseeds and closing the cluster. This continues until all the sequences are removed from the pool.

In the breaking phase (which happens during the growth phase), Swarm finds and breaks linked clusters which is closely related but does not belong in the same cluster. This linking can happen because Swarm is a single-linkage clustering method, and the molecular markers could be short and slowly evolving (Mahé et al., 2014) and therefore also very similar. Swarm uses the information about the internal structure of the clusters and the abundance values for all the sequences to break these chains of possible linked clusters. It is assumed that the most abundant sequences in a cluster have central positions and is surrounded by sequences that are less abundant. Swarm makes a graph representation of the OTU clusters where the most abundant sequences are peaks and linked clusters have paths between them. A path between highly abundant sequences is broken if the abundance value decrease and then increase again along this path. The centroid/representing sequence of an OTU is chosen to be the most abundant sequence of that cluster.

Because Swarm depends on continuous linking between sequences to cluster them together in the same OTU, it can occur clusters containing only one or two sequences (i.e., singletons or doubletons) that should be clustered together with more abundant clusters. This problem is solved by assuming the existence of a linking sequence between the low abundant OTU and the cluster it is added to. This is also solved by the microvariant approach. Microvariants are produced for all amplicons in clusters with an abundance of one or two. These microvariants are cross-checked with the microvariants from the high abundant amplicons, and the low abundant amplicons are linked with the high abundant amplicons if there is a match. This procedure is activated with the fastidious option, and reduces under-grouping of closely related amplicons (Mahé et al., 2015).

2.3 Taxonomic classification

A taxonomic classification of the centroid sequences from the four methods was performed using Kraken (Wood and Salzberg, 2014). This is not the most common method of classifying 16S reads. However, it ended up being the logical choice, because of the high cluster number produced by Dereplication and the high speed offered by Kraken to perform the classification. A short description of how the classification is performed, modified from the article by Wood and Salzberg (2014), follows.

Kraken use k-mers (short words with length k from the sequences) to search a database and classify sequences. All k-mers from a DNA sequence S are collected into a set ($K(S)$). These k-mers are mapped to the lowest common ancestor (LCA) taxon of all genomes that contain that k-mer. A classification tree is formed from the LCA taxa and their ancestors (in the taxonomy tree), which is a pruned subtree utilized to classify S. The nodes in the classification tree are associated with different taxa, and k-mers from $K(S)$ are mapped to these nodes. Each node is weighted according to the number of k-mers mapped to this particular node (representing a taxon). By calculating root-to-leaf (RTL) path scores, which is the sum of all node weights along the path, a maximum scoring RTL path (a classification path) can be found and S is classified to the label corresponding to its leaf. If there are multiple paths with the same maximum path score, the selected classification label is the LCA of those paths' leaves. Kraken will not classify sequences where none of the k-mers in $K(S)$ are found in any genome.

The database used by Kraken to perform query searches, needs to be pre-computed in order to obtain efficient implementation of the classification algorithm. A chosen library of genomic sequences is turned into a database of every distinct 31-mer in the library. When the unique k-mers are detected, each genomic sequence are processed to set stored LCA values of all k-mers in the sequence from the taxon that is associated with the sequence. For the classification performed in this thesis, the standard database included in Kraken were used. This database is based on the completed microbial genomes in the RefSeq database from the National Center for Biotechnology Information (NCBI). The database is something that can be changed by the user.

2.4 Comparison of methods

To get an overview of how and in what degree the clustering differs for the four methods, I performed several comparisons between them. To be able to better compare the clustering of the amplicons, the data set were clustered both in total (which is the most common procedure) and each sample separately. The comparisons were performed on both clustering approaches. Some of the methods used to compare the clustering methods and the composition of the output are described in this section.

2.4.1 BLAST

A sequence comparison was performed to get an overview of the similarities between the centroid sequences representing each cluster from each sample, in the sample by sample clustering. This comparison was executed using the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). BLAST uses heuristics to find similarities between a query sequence and a database of sequences. The algorithm does not guarantee the production of optimal alignments, but utilizes k-mer words to produce local alignments between sequences quickly (Madden, 2013). The k-mers are short words made from the query sequence, which are used to search for matches in the database. A database, from the centroid sequences, were created from a Fasta file, resulting in an indexed BLAST database which is faster to perform searches in (Madden, 2013). The output from the search, utilized in this study, is a percent identity score giving information about how similar two sequences are. megaBLAST from the Blast+ software was used to perform the comparisons (Camacho et al., 2009). megaBLAST is optimized for sequences that are very similar, and starts with a search for an exact match of 28 bases, then performs an attempt of extending this initial match into a full alignment.

2.4.2 ANOVA

An analysis of variance (ANOVA) was performed on the sample by sample clustering result to see if there were any effect of method on the mean cluster size (singletons were not included). ANOVA is, in general, a method to check if there is a difference between the within group variation and the between group variation. ANOVA is the simplest form of the F test, an extension of the t test, where more than two groups are compared (Lindman, 1992). A hypothesis is tested, where the null hypothesis states that there is no difference between groups, i.e. the variation within groups

and between groups are the same. Versus the alternative hypothesis that there is a difference between groups, i.e. the within group variation and the between group variation differ significantly.

When performing an analysis of variance we need to make some assumptions about the data; the observations in the groups must be normally distributed with mean μ_i and variance σ^2, and the variance of all groups are equal and the scores are independent of each other (Lindman, 1992). The analysis were performed in R by fitting a linear model, using the function *lm* with mean size as response and method as prediction factor.

2.4.3 Unifrac Distances and MDS

Multidimensional Scaling (MDS) was performed on the Vsearch total clustering result to get an impression of how similar the samples were in terms of OTU composition, and to see if there where grouping of any of the sampling categories. MDS is a method to visualize dissimilarities in a data set. The main idea is to find a lower-dimensional way of representing the data, in which the pairwise distances are as well preserved as possible (Hastie et al., 2009). The MDS analysis uses a distance measure, and generalized weighted UniFrac distances were calculated to measure distances between the samples in the honey bee data set. For the MDS analysis the R function *cmdscale* was used.

The R package GUniFrac (Chen, 2018) was used to calculate the distances. To calculate the generalized weighted UniFrac distances the function uses both the OTU table and a rooted tree computed based on the centroid sequences from the OTUs. The UniFrac distance takes into account the phylogenetic relationship in the data when calculating the distances (Chen et al., 2012). The generalized UniFrac distance, proposed by Chen et al. (2012), unify the original weighted and unweighted UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007). The weighted UniFrac distance use information about the species abundance and branch lengths, in the phylogenetic tree, are weighted with difference in abundance. The unweighted distance use only the information about present and absent species and the fraction of branch length unique to either community are counted. The generalized UniFrac distance combines these features and are designed to be robust and powerful in the detection of biological changes in microbiome composition (Chen et al., 2012). The distance measure have an α which can be varied between [0,1] to control the contribution from high-abundance branches, where $\alpha = 1$ means most emphasis on these branches. In this study the α was set to 0.5, and the distance is defined (when $\alpha = 0.5$), as (Chen et al., 2012):

$$d^{(0.5)} = \frac{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B}}$$

Where A and B are two microbiome communities, and n is the number of branches in a rooted phylogenetic tree. Further, b_i is the length of the branch i and descending from this branch the taxa proportions p_i^A and p_i^B for community A and B, respectively.

2.4.4 Phylogenetic Trees

Phylogenetic trees were produced to get insight in the phylogeny of the OTUs from the clustering methods. The trees give information about how similar sequences are regardless of the taxonomical classification. Several steps are involved in the making of a phylogenetic tree, the first step is to create a multiple alignment of the centroids. The multiple alignment was produced by the Muscle algorithm (Edgar, 2004) accessed through the microseq R package (Snipen and Liland, 2018). From the multiple alignment, pairwise distances between the centroid sequences was computed. These distances were calculated using the ape R package (Paradis and Schliep, 2018) (*dist.dna* function) with the evolutionary model proposed by Tamura and Nei (1993). The Tamura/Nei model does not assume that the base frequencies are equal and thus, estimate them from the data. The transition ($A \leftrightarrow G/C \leftrightarrow T$) and transversion (changes between purines and pyrimidines) rates are both presumed to be distinct. The distances are calculated based on the estimation of the number of nucleotide substitutions between sequences. A Neighbor-Joining (NJ) tree (Saitou and Nei, 1987; Studier and Keppler, 1988) is then inferred from the distance matrix, also using the ape R package. The NJ algorithm is agglomerative, and starting from a star shaped, unresolved tree the algorithm iterates these three steps (Criscuolo and Gascuel, 2008):

1. Join the taxon pair xy, with the smallest distance value, into a new node u
2. Estimate the length of the new external branches ux and uy
3. In the distance matrix; replace x and y by u, and estimate new distances between u and the remaining taxa

When the tree is completely resolved, the iteration stops.

In the resulting phylogenetic tree the nodes were colored according to the prevalence of the OTU represented. The prevalence says something about the presence of an OTU across samples. If the prevalence is 100 %, that means that the OTU is detected in all samples. This gives a graphical representation of potential core OTUs.

2.4.5 Alpha diversity

Alpha diversity measures the diversity of species/OTUs within a sample (Hugerth and Andersson, 2017). A simple, qualitative definition is that alpha diversity is high if a sample contains a high number of species/OTUs that is equally abundant, and low otherwise (Finotello et al., 2016). When all species/OTUs are equally abundant, the alpha diversity is at its maximal (for a given number of different species/OTUs). There are several different ways of measuring the alpha diversity, and a consensus method has not been found. In this study the Shannon entropy was used, which takes the total number of different species/OTUs and the species/OTU relative abundance into account when calculating the diversity (Finotello et al., 2016). The Shannon entropy is defined, mathematically as (Shannon, 1948):

$$H = -\sum_{i=1}^{S^{obs}} p_i \cdot ln(p_i)$$

Where S is the number of species/OTUs observed and p_i is the relative abundance of OTU $i = 1, .., S$.

The Alpha diversity was measured for each sample in all four OTU tables; Vsearch, Dada2, Swarm, and Dereplication. A mean value for each gut part in each month were calculated from the diversity measures. This was done to get insight in the differences between clustering methods based on diversity, and also to see if there were variations in diversity between months or gut parts. The R package vegan (Oksanen et al., 2019) was used to calculate the alpha diversity values.

Chapter 3

Results

3.1 Clustering sample by sample

In order to see the effect of the various clustering methods, each sample was initially analyzed on its own. This gives one set of results for each method, making comparisons more stable. There are 460 samples in total and the reads in each sample were clustered using four different methods; Vsearch, Dada2, Swarm, and Dereplication.

3.1.1 Cluster Number and Size

Clusters produced for each of the methods were counted. The number of clusters say something about the resolution of the method, i.e. how many clusters the reads are divided into. In a typical clustering the result will consist of a smaller number of large clusters and a larger number of small clusters. The smallest cluster possible has only one member, and are referred to as singletons. It is quite common to discard these singleton clusters, presuming they are sequencing error artifacts (Behnke et al., 2011). In the comparisons the singletons are both included or excluded, and this is specified in each case. The cluster number were computed relative to the Vsearch cluster number to get an impression of how the other, newer methods (Dada2, Swarm) compares to the more traditional 97 % identity clustering (Vsearch) (figure 3.1). Swarm, Dada2, and Dereplication all have more clusters than Vsearch. Swarm is the method that differ the least from Vsearch. Dada2 have quite high cluster numbers compared to Vsearch, and are actually more similar to the Dereplicated numbers. It is as expected that Dereplication have very large cluster numbers compared to Vsearch, because errors in the sequencing lead to many unique sequences.

Cluster number affects cluster size, because the input read count is the same across methods and more clusters will further divide the reads among them. Cluster size for all samples are displayed as boxplots in figure 3.2, the size is represented as read count. All sizes for all samples were registered, in addition to the mean size for each sample. Vsearch and Swarm vary a lot in mean size, and are the methods with the largest spread. Dereplicated have, as expected, the smallest cluster sizes.

In addition to the boxplots, a linear model was fitted and an ANOVA was performed with the mean size as response and the method as prediction factor. This analysis was executed to examine if the mean cluster sizes of the four methods were significantly different. The model used Vsearch as reference factor, and the result

yielded negative estimates of the mean values, for the other methods, with significant p-values. This indicates that the differences between mean size in the methods really is different and that Dada2, Swarm, and Dereplication have smaller size than Vsearch. The ANOVA output gave a p-value below 0.001, meaning that the within mean size variation for each method is different from the between method mean size variation.

Table 3.1: A table containing the mean and standard deviation for the per sample singleton count for each method.

	Vsearch	Dada2	Swarm	Dereplication
Mean	145	1.17	3456	15 944
Std.	115	2.64	2507	13 085

The methods produce a varying amount of singletons for the sample by sample clustering. The mean, per sample singleton count for all four methods were computed (table 3.1) and Dada2, unsurprisingly, have the lowest amount of singletons with a mean count of 1.17. Dereplication have the highest count with 15 944 singletons, but this is expected as there is no clustering step for these sequences. Swarm have a very high number of singletons compared to Vsearch and Dada, even though the fastidious option, which is meant to prevent the production of a large amount of singletons, was used during the clustering step.

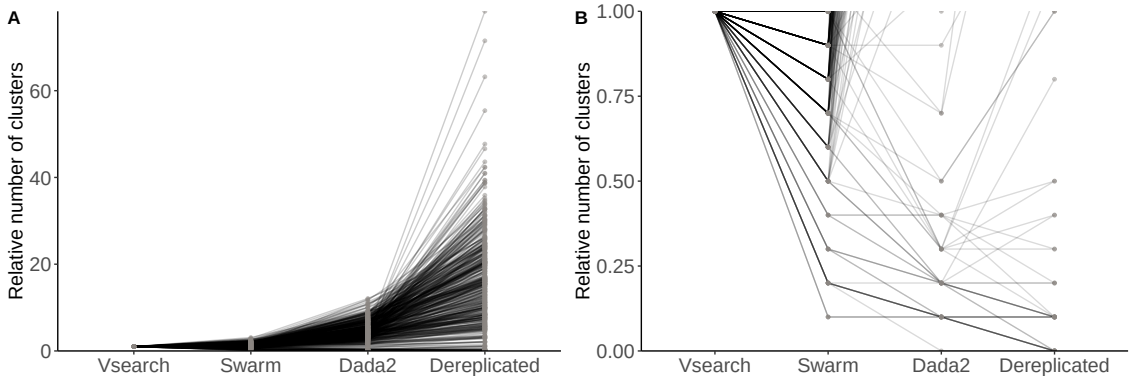


Figure 3.1: Number of clusters relative to the Vsearch cluster number for each method, each curve corresponds to a sample. The methods displayed are Vsearch, Dereplicated, Dada2, and Swarm. Singletons are not included in the cluster number for any of the methods. (A) All curves are displayed (B) Only the interval between zero and one on the y-axis is displayed. These are the samples that have a lower number of clusters compared to Vsearch.

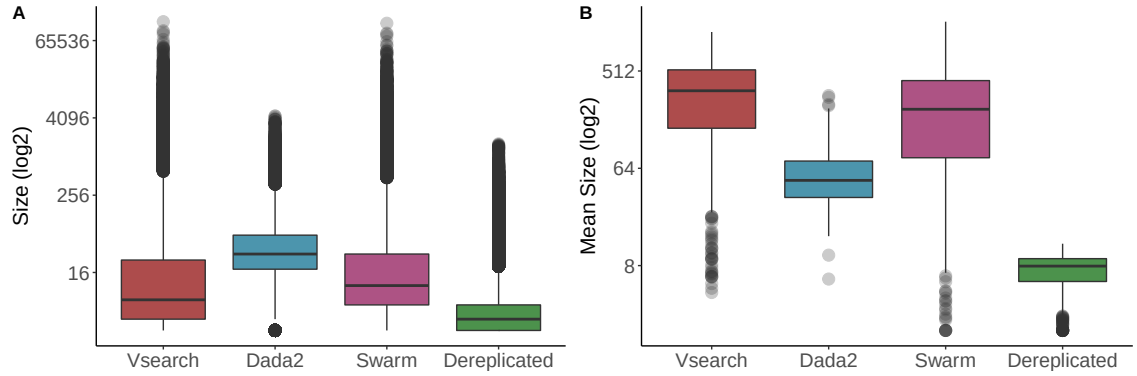


Figure 3.2: Box plots representing the size of the clusters, in read counts, for the different methods when samples are clustered separately. Singletons are not included and the sizes are log_2 transformed. (A) All cluster sizes from all samples are included (B) Only the mean cluster size from each sample is included.

3.1.2 Difference between clusters

A pairwise comparison between the four clustering methods were performed to determine similarity between the sequences representing the OTUs. A BLAST search was performed with the centroids from one method as query and the centroids from another method as database. The search was done sample by sample, and a mean percent identity was computed for each sample in each pairwise comparison. The final result of the comparison is displayed in figure 3.3, where the percent identity represents the similarity between the centroids from the different methods. The queries without a hit in the database was included as 0 % identity to account for the difference in cluster counts created by the different methods. Each unique OTU from both the query and the database were only registered once in the computation of the mean percent identity for each sample.

Swarm and Vsearch seem to be the most similar methods, with the highest mean percent identity value. Dada2 have a higher similarity to Dereplication than both Vsearch and Swarm, perhaps indicating that Dada2 produce a higher cluster number.

3.2 Clustering the entire data set

In order to find the core microbiota, it seems more natural to cluster sequences by considering all samples at once. For this exercise, the same four methods were used:

1. Standard 97% identity clustering, Vsearch
2. Dada2
3. Swarm
4. Full Dereplication

For the methods Vsearch, Swarm, and Dereplication the samples are now pooled before clustering all reads in all samples in total. Hence, creating clusters containing amplicons across samples, not limited to containing only the reads from one sample.

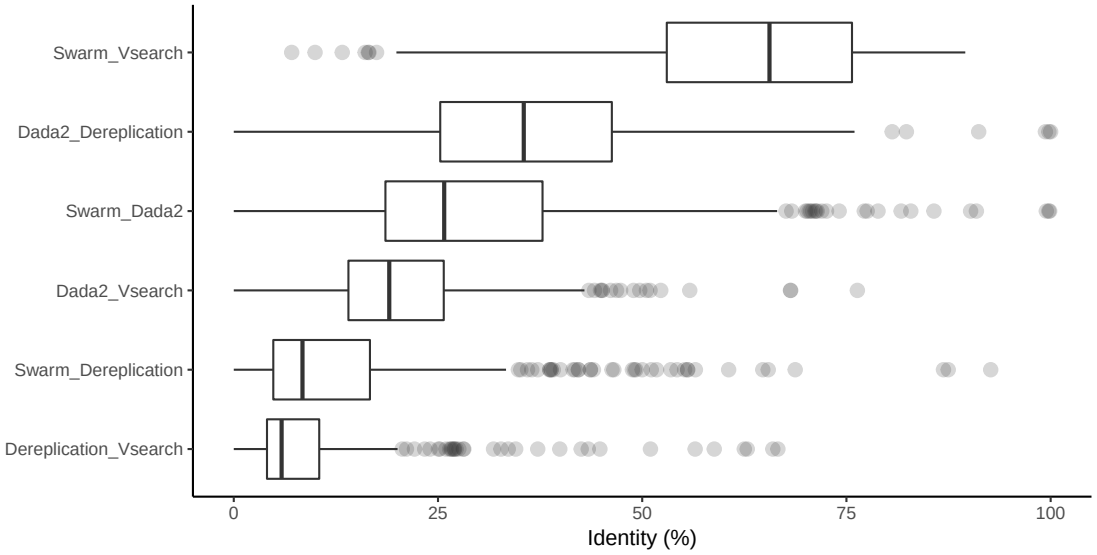


Figure 3.3: Boxplots representing the percent identity of centroids between pairwise compared samples. The centroids in each sample are BLASTed against the centroids in the corresponding sample from another method. The labels on the y-axis represents the methods compared, the left as query and the right as database. Singletons are not included.

The Dada2 clustering algorithm does not pool the samples (unless it is specified) before performing the clustering, and join the results of a sample by sample clustering into one big table after the sequences are denoised/clustered. Hence, the only difference between the sample by sample clustering and the total clustering, regarding the Dada2 results, is that the table is divided into individual sample tables containing read counts of single samples for the former. This type of clustering is possible because Dada2 claims to find exact sequence variants, these variants should be the same across all samples, if they have the same biological sequences. Singletons are not included for any of the methods in the total data set clustering.

3.2.1 Cluster tables

From the four clustering methods there are four resulting OTU tables. The tables vary a lot in size, with the Vsearch table ending up at around 500 OTUs as the smallest and the Dereplication table having nearly 700 000 OTUs, hence being the biggest one. Dada2 and Swarm ended up at around 7000 and 1100 OTUs, respectively. All tables have 460 samples, with abundance counts representing each OTU in each sample. The Dada2 table is missing counts for one sample (Jul_K3_Rec_5), due to failure to merge the reads in this sample. The sample is still included in all analyses, with read counts in the Dada2 table set to 0.

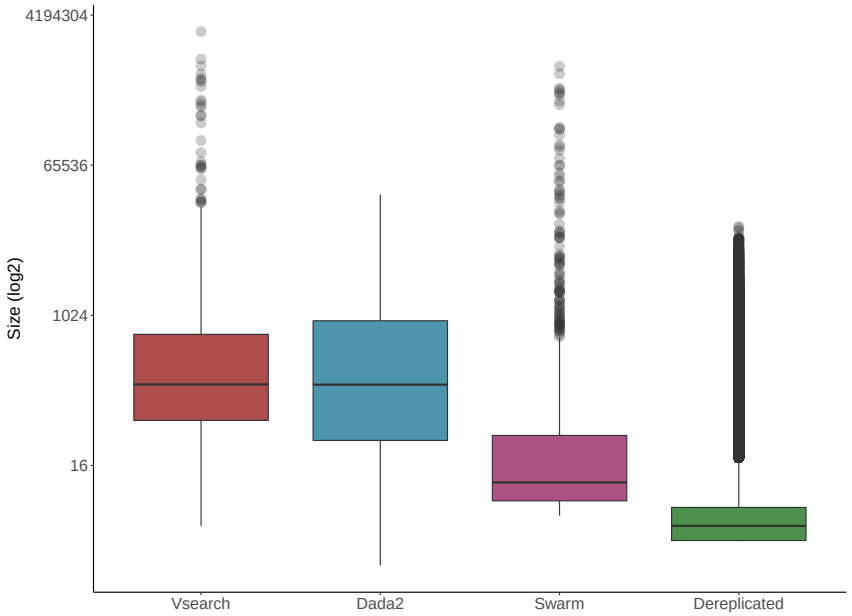


Figure 3.4: Size of clusters (read counts) for the entire data set clustered with the different methods. The sizes are log_2 transformed.

Cluster size

Cluster size were counted for the total clustering, similar to the sample by sample clustering, however the cluster size is the read count, per OTU, across all samples in the OTU table. The cluster sizes are displayed in figure 3.4. Vsearch and Swarm both produce some very big clusters and a larger amount of smaller clusters. Dada2 is much less variable in cluster size, and does not produce the same huge clusters as Vsearch and Swarm does. The Dereplication method contains mainly small clusters, however there is a large group of bigger sized clusters as well.

Rarefaction curves

Inspired by ecological studies of species richness with rarefaction curves, similar curves were computed for all four OTU tables and displayed in figure 3.1. When species richness is measured it can be problematic because when more individuals are sampled, more species/OTUs are recorded (Bunge and Fitzpatrick, 1993). A sampling curve will, at first, rise rapidly when there are many new taxa/OTUs observed in the first samples. For later samples, when increasingly rare taxa/OTUs are added, the curve will rise much more slowly (Gotelli and Colwell, 2001). Eventually if the sampling is extensive enough, and the spatial limits are not too big, the curve will reach an asymptote and no more taxa will be added.

The curves in this study were computed by plotting the cumulative number of new OTUs observed as a function of sample number, when samples were drawn at random. The cumulative number plotted is the resulting mean after 50 permutations. The curve from the Vsearch OTU table (figure 3.5A) is rising fastest of all the curves, and also seem to reach the asymptote at the top of the curve. Also the Swarm curve (figure 3.5C) rise pretty fast and grow more slowly at the end. Dada2 (figure 3.5B) on the other hand have a curve that rise slowly and does not seem to reach the asymptotic state. The curve for the Dereplication OTU table (figure

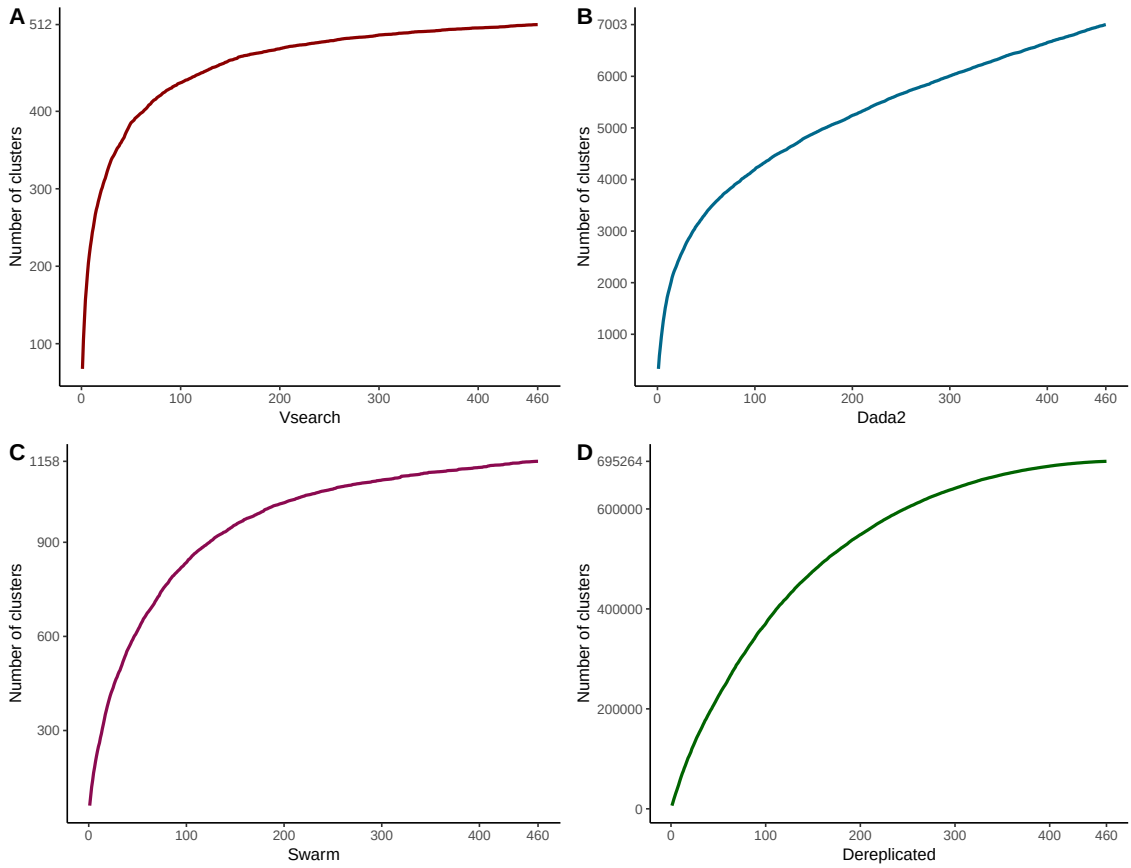


Figure 3.5: The cumulative number of clusters when samples are drawn at random. Only the unique clusters for each sample are added to the count. The graphs are for the full data set clustered with the different methods. The values are mean values computed after 50 permutations. (A) The data set is clustered with the Vsearch method. The number of clusters formed with this method is 512. (B) The data set is clustered with the Dada2 method. The number of clusters formed with this method is 7003. (C) The data set is clustered with the Swarm method. The number of clusters formed with this method is 1158. (D) The data set is clustered with the Dereplication method. The number of clusters formed with this method is 695 264.

3.5D) rise slowly, however it shows signs of reaching the steady growing asymptotic state at the end of the curve. If a curve fails to reach the asymptotic stage, and develop into a linear curve, it can be a sign of noise in the data. This failure to reach an asymptotic curve is what we observe for Dada2.

3.2.2 Core microbiota

In addition to the uncovering of how the clustering methods differ in structural comparisons, it was of interest to detect the core microbiota in the honey bee gut samples. And further, to get insight in the differences between methods with regards to the detection of a core microbiota.

Sample Prevalence

To get an overview of how many samples each cluster from the four methods contained, the per cluster sample count were displayed as bar graphs (figure 3.6). These figures made it possible to quickly observe if there were potential core OTUs present in all or almost all samples.

Vsearch and Swarm are the only methods with OTUs that have reads present in all samples (i.e. core microbiota). Dada2 is relatively far from any core OTUs, with clusters merely present in about half of the samples. A rather surprising result is that Dereplication have clusters present in more samples than Dada2. That is, unique sequences present in more than 300 samples in the Dereplication result, which is not present in the results presented by Dada2.

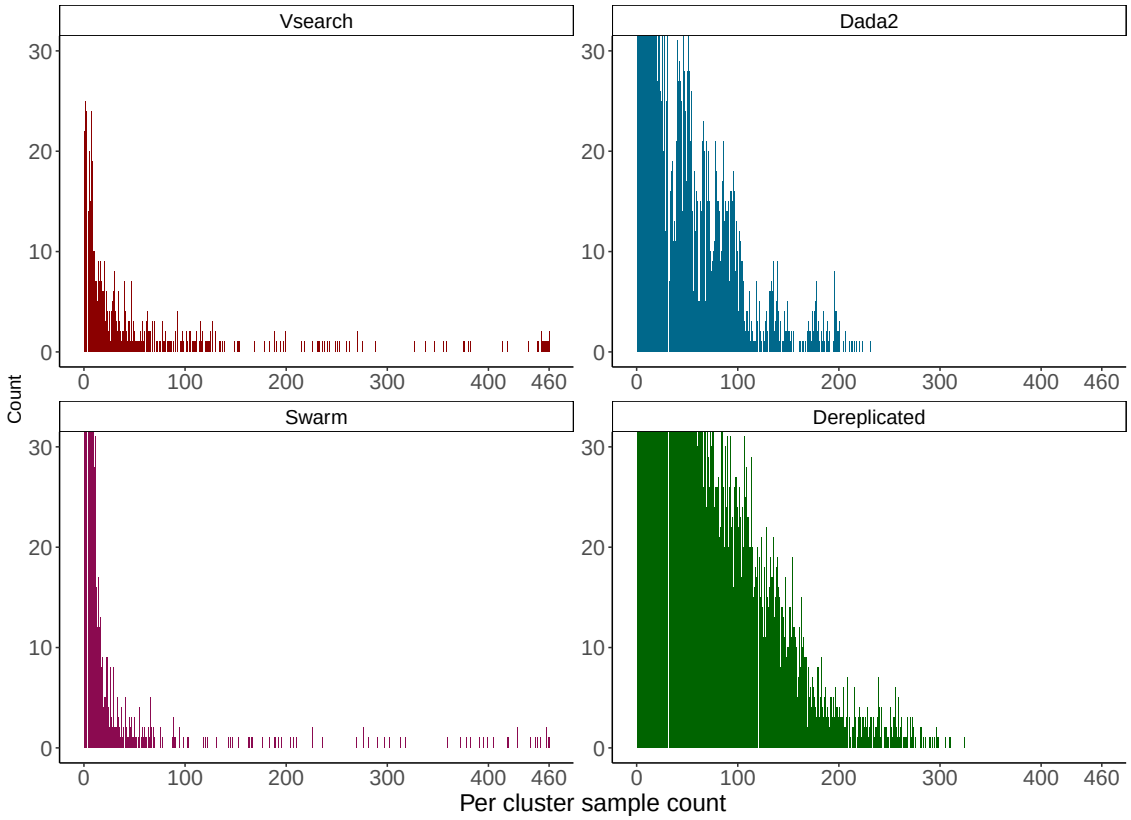


Figure 3.6: A set of histograms displaying the per cluster sample count for the full data set clustered with the different methods. The y-axis is cut at 30 to display the counts for the high per cluster sample counts better. The maximum number of samples a cluster can appear in is 460.

3.2.3 Grouping of samples

In the investigation of the core microbiota, it was also examined if there were any grouping of samples based on gut part, sampling month or sampling hive. If there were grouping of samples this could also mean that the microbiota composition in the different sampling categories are different.

Multidimensional scaling

An MDS analysis were performed from the Vsearch OTU table to see if there were grouping of samples based on gut part or hive. To create the MDS, weighted UniFrac distances were used as measure of distance between samples. The results show no evidence of any grouping of samples based on which hive they were sampled from. There is, however, an indication of grouping in regards to which gut part the sample is from. Crop, ileum, and rectum are the samples with best separation, while the midgut samples spread more across the other samples. A plot colored by month was also made. However, there was no evidence of any grouping based on month, hence it is not included.

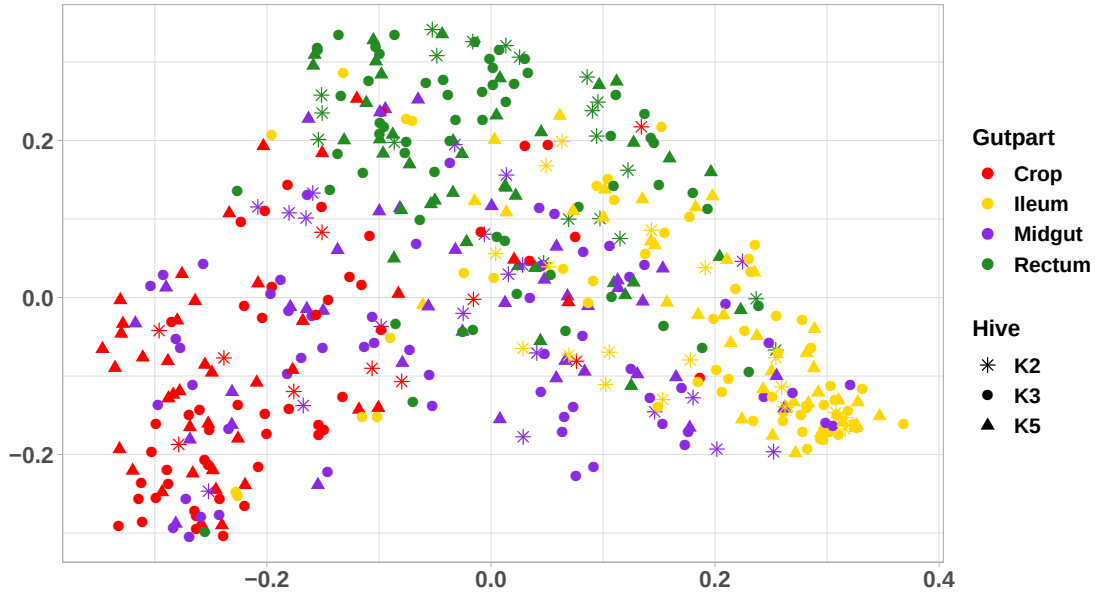


Figure 3.7: MDS based on the OTU-table for the Vsearch clustering of the full data set. The distances used to perform the analysis are weighted UniFrac distances.

Measure diversity

For further analysis of grouping and effect of sampling categories, a measurement of diversity were calculated from the OTU tables. Alpha diversity was measured as Shannon Entropy, and the OTU tables were divided by sampling month and gut part. The overall diversity values differ between methods, however this is expected as the OTU tables contain different numbers of OTUs (figure 3.8). Diversity values within each gut part are relatively stable for all methods, with no major increase or decrease. The Vsearch and Swarm diversities are nearly identical for all gut parts and months. In the crop gut part the diversity increase for Vsearch and Swarm in August and November, however, for Dada2 the diversity decrease. This ambiguous result will lead to opposite conclusions for the change in diversity of bacteria in the fall for crop samples, depending on method.

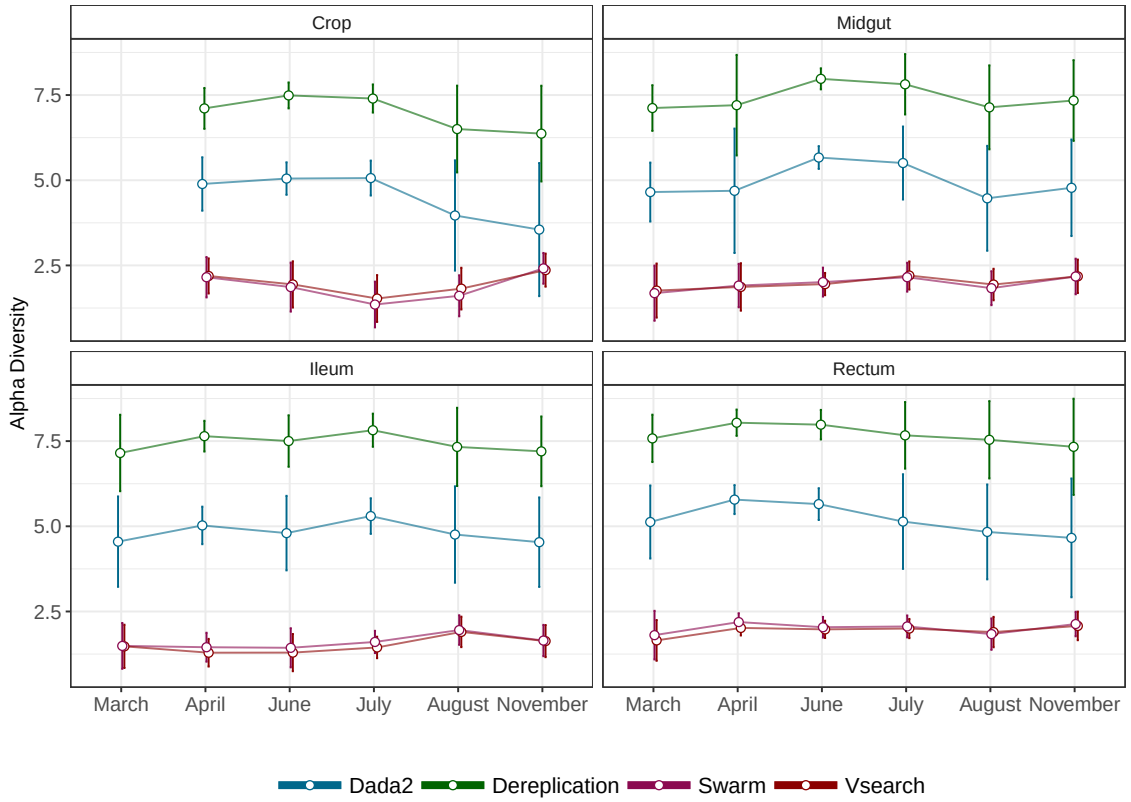


Figure 3.8: Alpha Diversity measured in Shannon entropy. The diversity is calculated from the OTU tables for the clustering of the entire data set. Points are mean alpha diversity and the error bars represent the standard deviation. The Diversity measures are grouped by gut part.

Gut part sample prevalence

With the results from the MDS analysis indicating a grouping of samples based on gut part, I decided to perform a per sample cluster count on the gut part samples from the OTU tables. All four OTU tables were separated into four smaller tables, one for each gut part, and all samples containing less than 500 reads were removed. Samples were counted in the same manner as for the total data set, resulting in four figures displaying the per cluster sample count for each gut part in each method (figures 3.9, 3.10, 3.11 & 3.12).

The results from Vsearch and Swarm are similar to the results from the count of the total OTU table, i.e. there are OTUs in all gut parts that are present across all samples. There was an increase in the number of core clusters, however this was expected when there is a decrease in possible samples to be present in. Dada2 does not have any improvement in the detection of core OTUs when the data is divided by gut part. Similar to the count for the total data set (figure 3.6), Dereplication have clusters with a higher sample count compared to Dada2 in this result as well. This is, again, very unexpected as the Dereplication contains the unique sequences. When the exact same sequence is present in that many samples, in Dereplication, there is little reason to believe that the sequence is erroneous. Thereby leading to the assumption that the sample by sample clustering procedure performed by Dada2, give a false estimate of the prevalence of clusters.

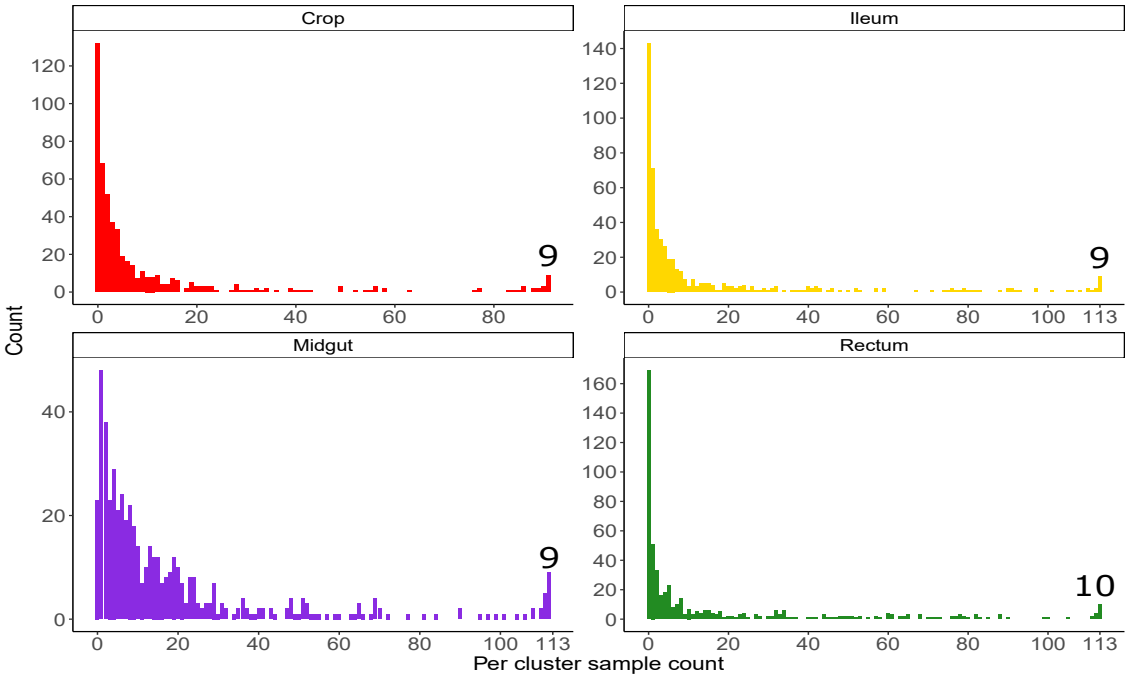


Figure 3.9: Vsearch
The number of samples per cluster plotted as bar graphs for the total data set. The number above the bar in each plot represents the number of clusters with 100% prevalence in samples.

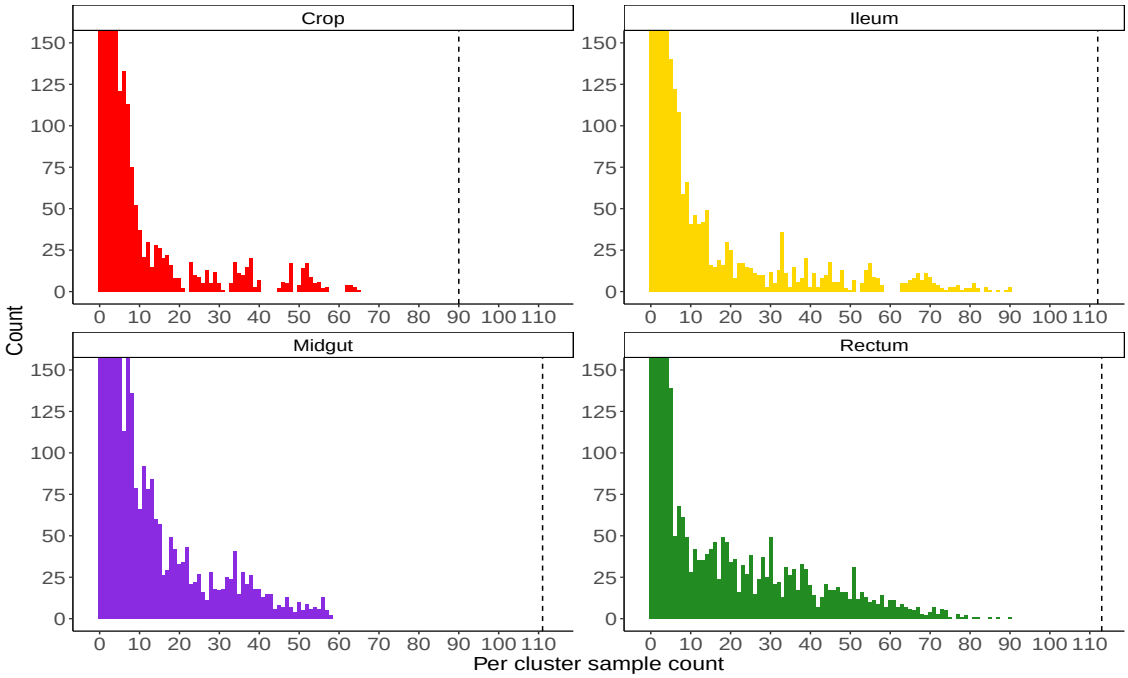


Figure 3.10: Dada2
The number of samples per cluster plotted as bar graphs for the total data set. A dashed line is drawn to indicate where 100% sample prevalence in a cluster would be. The y-axis is cut at 150 to display the lower bars on the right better.

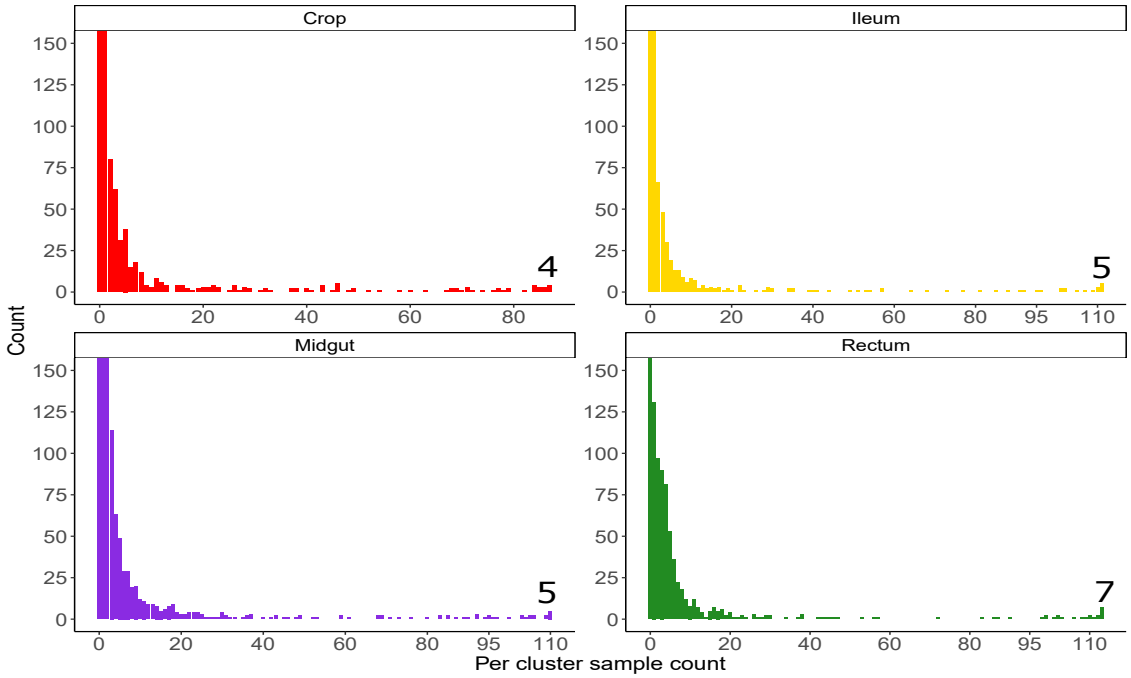


Figure 3.11: Swarm
The number of samples per cluster plotted as bar graphs for the total data set. The number above the bar in each plot represents the number of clusters with 100% sample prevalence. The y-axis is cut at 150 to display the lower bars on the right better.

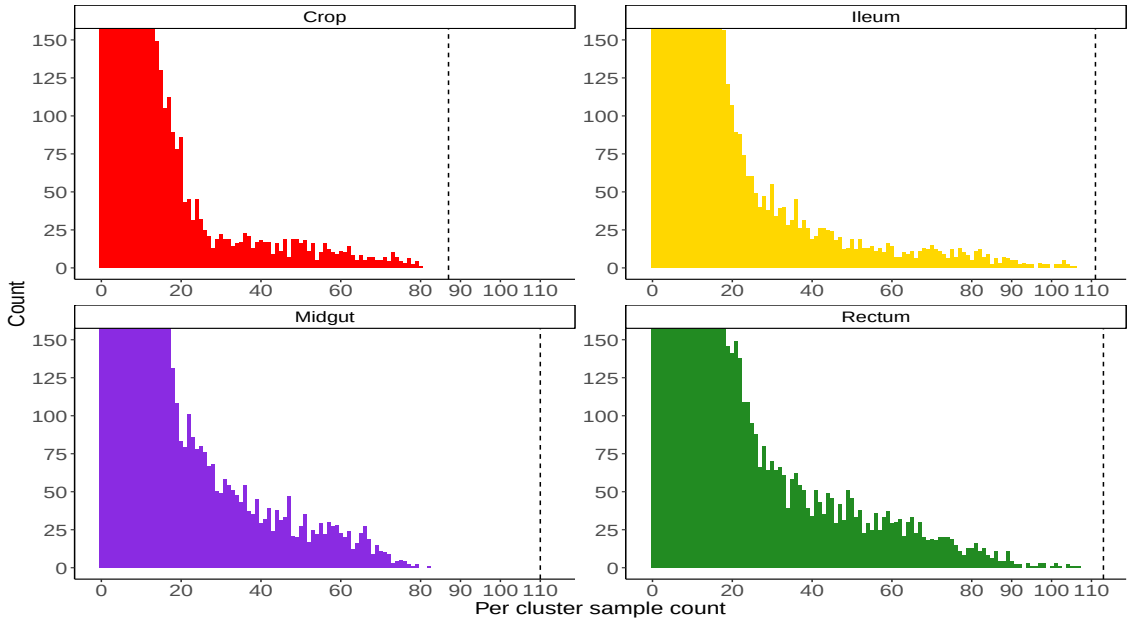


Figure 3.12: Dereplication
The number of samples per cluster plotted as bar graphs for the total data set. A dashed line is drawn to indicate where 100% prevalence of samples in a cluster would be. The y-axis is cut at 150 to display the lower bars on the right better.

3.2.4 Phylogentic trees

Centroids from all methods were taxonomically classified using the Kraken algorithm. Genus level classification were used for all analyses in this study. The proportion of centroids classified to the genus level for the methods were; Vsearch 76.5%, Swarm 87.3%, Dada2 92.4%, Dereplication 98.7%.

Phylogenetic trees were produced from the centroids in the OTU tables. These trees were produced to get further insight in the core microbiota, and to look at the phylogenetic relationship between the centroids. The phylogenetic relationship give information about the sequence similarity, regardless of the previous taxonomic classification. The nodes in the trees are labeled with genera and colored by prevalence in samples.

Initially, a tree for the total results was produced to get an overview of the core OTUs detected in the per cluster sample count (figure 3.6). Based on this previous detection of core OTUs in the data, a decision was made to produce only the Vsearch total prevalence tree (figure 3.13). It was assumed that a Swarm tree would be similar to that from Vsearch, regarding the core OTUs.
Prior to the calculation of the phylogenetic tree, the OTU table went through some filtration steps. In each sample, only the clusters contained within the top 99% of reads were included. OTUs with a prevalence below 10% were removed from the analysis, in order to reduce the number of OTUs in the phylogenetic tree.

The resulting Vsearch tree (figure 3.13) display three OTUs with very high prevalence from the genera *Lactobacillus*, *Gilliamella* and *Snodgrassella*. Within these genera there are nodes with lower prevalence values as well; hence, there are specific OTUs from the three genera that are present in all or almost all samples.

Further, because core OTUs was detected in all gut parts, a prevalence tree for each of the gut parts was produced with the same procedure as for the total Vsearch tree. The trees were produced to see if there was any differences in the prevalent genera compared to the total phylogenetic tree. There were only minor differences in the highly prevalent genera, and therefore the gut part trees are not included in this result section. Only the crop-tree was noticeably different in core taxa between the total and the four gut part-trees. A table is included with the genera that belong to the OTUs with a prevalence above 90% (table 3.2), i.e. nodes that are colored pink in the four trees. The phylogenetic trees belonging to each gut part can be found in the appendix (figure A.3-A.6).

Table 3.2: OTUs with a prevalence above 90% in each gut part. The genera are used as label.

Crop	Midgut	Ileum	Rectum
Pseudomonas	*Pseudomonas*	*Snodgrassella*	*Lactobacillus*
Sphingomonas	*Snodgrassella*	*Gilliamella*	*Lactobacillus*
Pseudoalteromonas	*Gilliamella*	*Lactobacillus*	*Lactobacillus*
Gilliamella	*Parasaccharibacter*	-	*Snodgrassella*
Parasaccharibacter	*Lactobacillus*	-	*Gilliamella*

Dada2 did not have any occurrences of core OTUs in the previous results, therefore it was of interest to examine if a merging of OTUs belonging to the same genus in the OTU table would lead to more prevalent OTUs. For each unique genus, all reads belonging to OTUs classified to this genus were added, resulting in one read count for each unique genus for each sample. This procedure reduced the OTU table to the unique genera, and with the merged read counts the probability of detecting core genera increase. A sequence representing each genera is required to produce a phylogenetic tree. Here, the centroid sequence belonging to the most abundant OTU in each genus was chosen as the representative sequence for the entire genus. The tree representing Dada2 is displayed in figure 3.14, this tree is very different from the trees produced for the other tree methods (figure 3.15 and figure A.1, A.2 in the appendix). Vsearch, Swarm and Dereplication produced genus trees with almost all nodes colored pink, hence a large portion of the genera have a high prevalence. The Dereplication genus tree (figure 3.15) is included as an illustration of this, the other two are fairly similar and can be found in the appendix (figure A.1, A.2). The Dada2 tree barely have three genera with very high prevalence, and these are the same genera found in the Vsearch original tree (figure 3.13). These genera have the following prevalence in the Dada genus tree; *Lactobacillus* 93%, *Gilliamella* 95%, *Snodgrassella* 86.9%.

A genus count, from the OTU tables, was performed to display the differences in the number of OTUs each clustering method produce for the top 10 most prevalent genera according to the Vsearch data (table 3.3). This table illustrate how many clusters that are merged for each genus in each method. Dada2 produce 305 OTUs for the genus *Frischella*, however when all these are merged the prevalence only reach 60% which is very different from all the other methods which ended up with approximately 98%.

Table 3.3: OTU counts for the 10 most prevalent genera according to the Vsearch method. The count is from the total OTU table for each method.

Genus	Vsearch	Dada2	Swarm	Dereplication
Gilliamella	45	1090	107	180 706
Snodgrassella	13	466	33	56 054
Lactobacillus	35	1324	213	144 008
Bifidobacterium	15	157	19	23 163
Frischella	20	305	27	38 465
Komagataeibacter	6	94	26	37 343
Parasaccharibacter	6	205	27	75 335
Bartonella	10	129	40	47 001
Pseudomonas	16	296	46	17 717
Acidisphaera	6	158	30	38 454

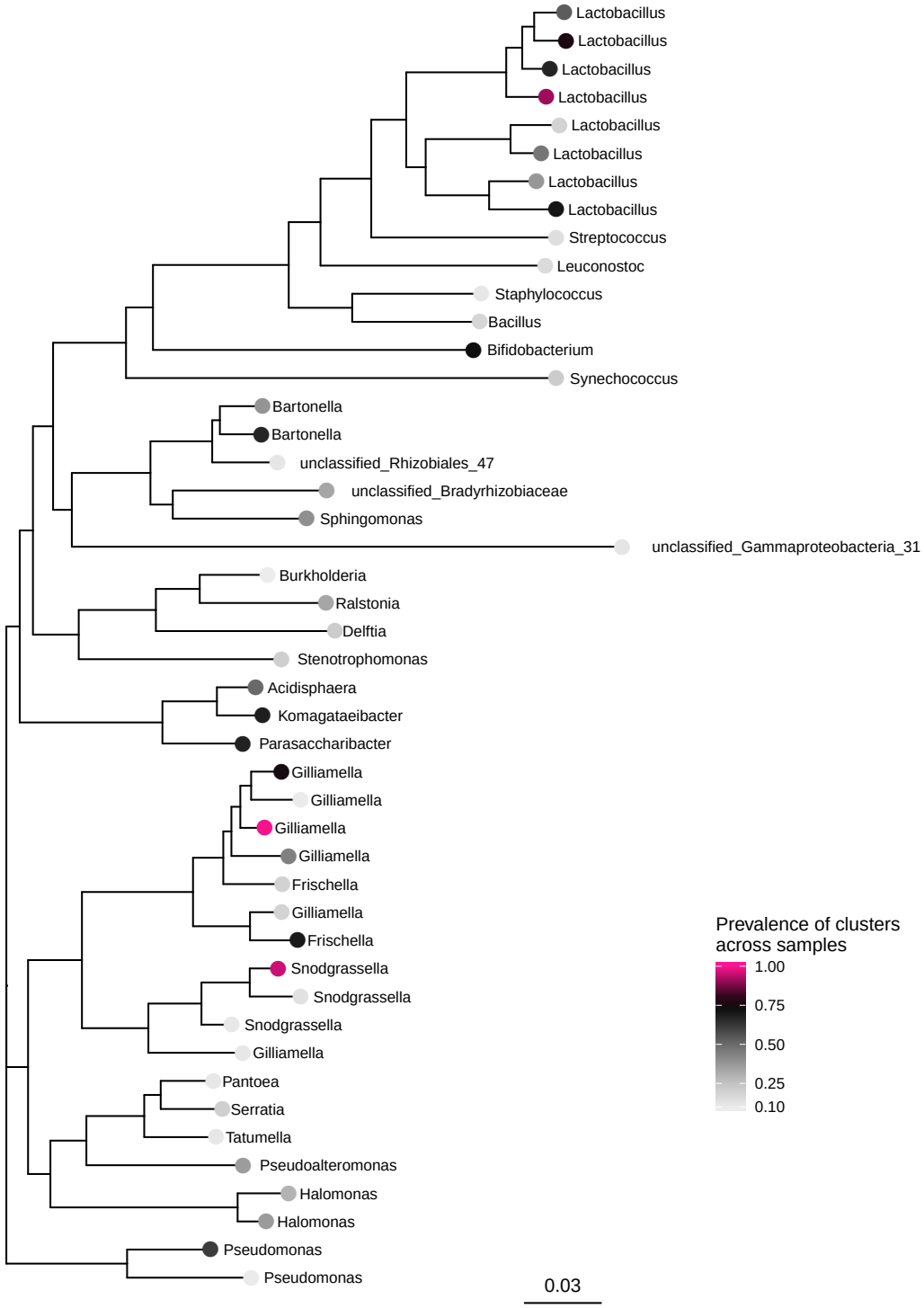


Figure 3.13: Prevalence tree for the Vsearch OTU table. The colored tip represents the prevalence of the cluster across all samples. The leaves are labeled by genera.

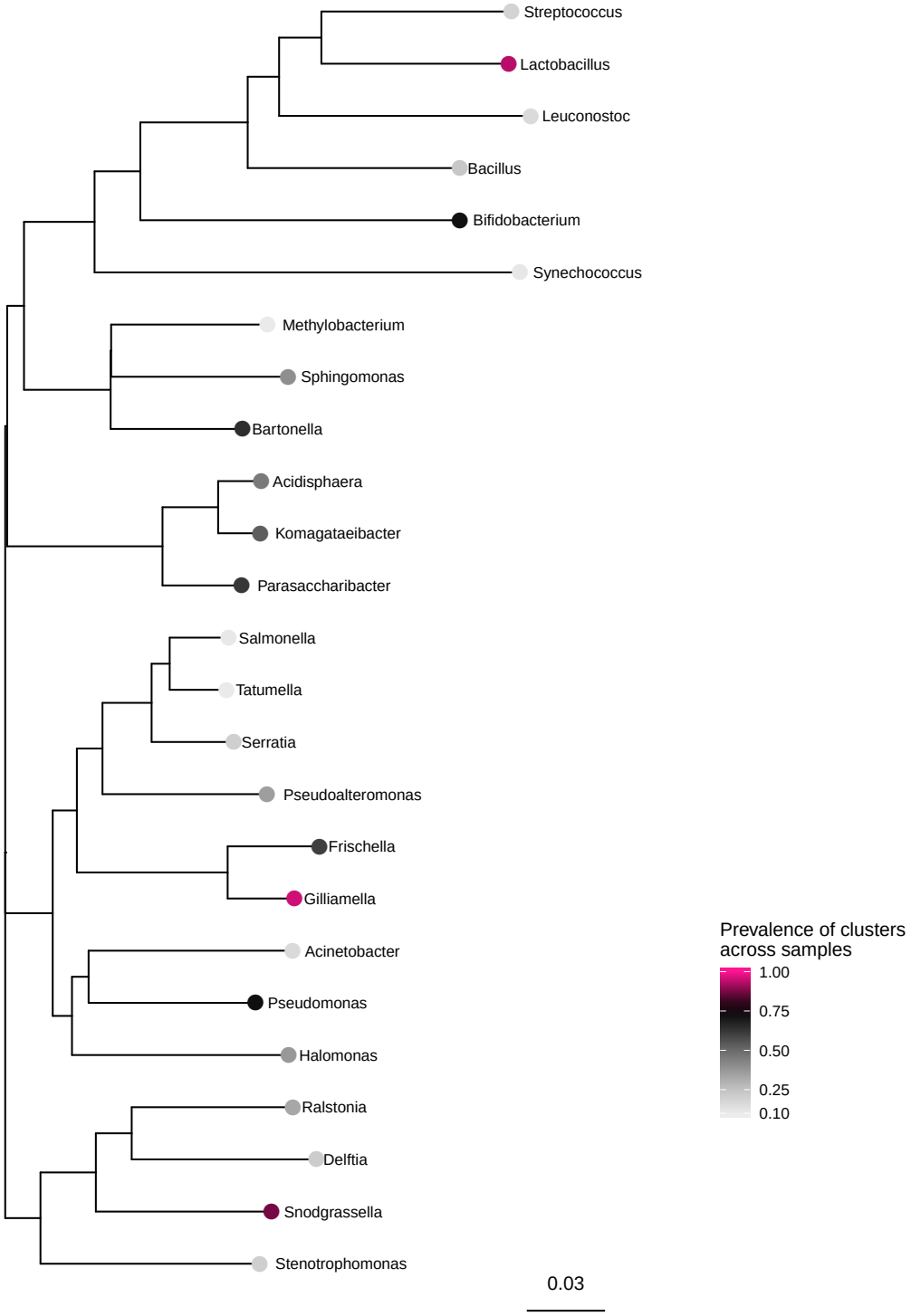


Figure 3.14: Genus prevalence tree for the Dada2 OTU table. The nodes are colored according to prevalence across samples in the merged genus table. Genera with a prevalence above 10% are included.

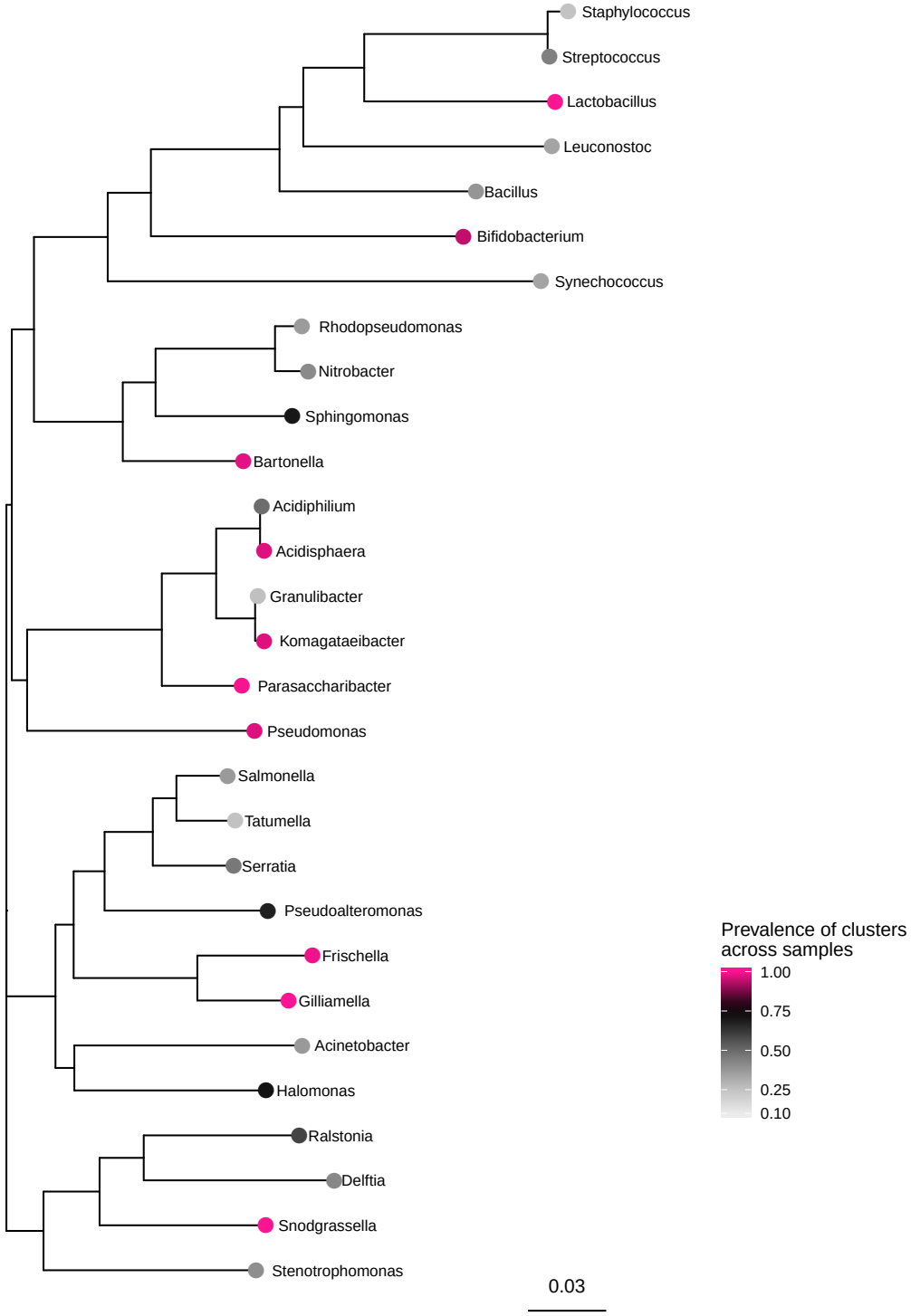


Figure 3.15: Genus prevalence tree from the Dereplication OTU table. The nodes are colored according to prevalence across samples in the merged genus table. The genera with a prevalence below 25% are not included.

Chapter 4

Discussion

4.1 Similarities in sample by sample clustering

In this thesis four main methods was used to cluster amplicon reads. In general they can be described as: clustering of unique reads, traditional identity clustering, a clustering algorithm based on single-linkage sequence differences, and a denoising algorithm made to infer exact sequence variants.

Dereplication is the most simple of the four, and is not really clustering. When reads are dereplicated each unique sequence are identified and the abundance is counted. Dereplication is the equivalent of sequence similarity clustering with an identity of 100%.
Vsearch is one of the most common methods of performing amplicon read clustering. Dereplicated reads are clustered based on sequence similarity with a fixed identity threshold, often set to 97%.
Swarm cluster the dereplicated sequences utilizing a graph based algorithm. A graph is produced from the dereplicated sequences, where each sequence is a node. Nodes that have a difference of 1 base are connected with edges, and all nodes connected in the same sub-graph result in a cluster.

Dada2 stand out from these three first methods, and cluster the sequences in a very different manner. The biggest differences is that Dada2 merge the forward and reverse reads after they have been clustered separately, and the algorithm use an error model when clustering the sequences trying to model errors in the data caused by sequencing. With these four methods the aim was to examine the differences in their cluster composition and get insight in the effect of method regarding the detection of core microbiota.

Samples were clustered separately at first to get a better insight in the similarities and differences of clustering between the four methods; Vsearch, Dada2, Swarm, and Dereplication. As one of the main aims of this study was to see how clustering methods impact the detection of core microbiota, it was of interest to examine cluster number and size as they are both factors that can impact this. Considering that all methods are given the same amount of reads as input, a larger number of clusters will affect the size in that the reads are spread across a larger number of groups.

Both cluster numbers and size differ greatly across methods. In the cluster numbers figures (figure 3.1) it is clear that Dada2, Swarm and Dereplication produce a larger amount of clusters than Vsearch in general. The large amount of clusters produced for the Dereplication method is highly anticipated as this method only

cluster the unique sequences into "OTUs". The majority of the samples have cluster numbers larger than Vsearch for Dada2, Swarm, and Dereplication.

Table 3.1 shows that the number of singleton clusters (clusters containing 1 read) varies a lot between the methods, but also between samples, considering standard deviations are in general of the same magnitude as the mean. Singleton clusters are commonly discarded from downstream analyses, because they are assumed to be erroneous sequences. It is worth noting that the Dada2 method produces extremely few singletons. This is inherent in the method as the clustering algorithm does not allow for singleton clusters to form their own partition (Callahan et al., 2016). The singletons being observed are explained in an answer to an issue filed on the Dada2 GitHub page as (https://github.com/benjjneb/dada2/issues/92, read: 06.07.2019):

> Singletons cannot appear in the raw output of the dada(...) function, but can appear after merging, if for example the reverse read of one member of a doubleton was misassigned due to low quality. This is relatively rare, but possible.

The merging procedure is poorly described in the article by Callahan et al. (2016), the only information provided is that the forward and reverse reads need to be in the same order when read in to the dereplication function. In the output from R it is difficult to understand what happened during the merging of the forward and reverse reads. For each sample, a table containing information about the merged denoised reads are output, and for each merged sequence an index for the forward and reverse merged denoised reads are listed. As far as I could understand, the indices indicates which denoised sequences are merged, but the same index is listed several times in both the forward and the reverse lists. This makes it seem like the same forward sequence is merged several times with different reverse sequences, and vice versa. After merging, the abundances of the denoised sequences have changed decreasingly compared to before merging. This indicates that Dada2 find several matching denoised sequences, and thus divide the reads across more clusters.

As mentioned in the Methods section, Dada2 merge the sequences after denoising to avoid disruption in the empirical relationship between the sequence quality score and the error rates. With the explanation of the merging from this data set and no further description given as to how the merging procedure is performed from Dada2, it is hard to understand what actually happens to the reads when they are merged. Other methods, like Vsearch, Swarm, and e.g. UNOISE2 (another denoising algorithm) (Edgar, 2016), utilize the information about which of the forward and reverse reads that are pairs given by the paired end sequencing, and merge the sequences before any clustering.

These observations agree with the explanation for the GitHub issue, but it does not guarantee that the only consequence is the inference of a few singletons.

The relative number of singletons, for the methods, may seem high, but will in general be larger when samples are small, due to the more shallow sequencing. Thus, by considering all samples together, the fraction of singletons will in general decrease, assuming overlap in composition between most samples.

The large amount of singletons produced by Swarm may influence the cluster number relative to Vsearch, these are discarded in the analysis (figure 3.1) and may be part of doubletons (cluster containing two reads) and other small clusters in the

Vsearch results. Swarm is the method with the largest amount of clusters with a lower relative count compared to Vsearch, however it is also the method that is most similar to Vsearch in regards to cluster number.

The cluster size figure (figure 3.2) also display the similarity between Vsearch and Swarm, both methods seem to have a similar cluster size composition. In the boxplot that present all cluster sizes (figure 3.2A) both Vsearch and Swarm produce a decent amount of very large clusters, while the main portion of clusters have a smaller size. Dada2 is much less variable in cluster size, and also have a lower overall mean cluster size value than Vsearch and Swarm. This may indicate that Dada2 produce a higher number of clusters with a more even size.

4.1.1 Centroid similarity

It can be difficult to compare centroids between clustering methods because the reads chosen as the centroids are not necessarily the same exact sequences. A factor that can lead to differences in the sequences that become centroids, is large variation in cluster numbers across methods. The most stringent way of doing the comparison is to demand 100% similarity between the centroid sequences, but then two clusters can consist of nearly the same sequences and have slightly different centroids, and thus be discarded as dissimilar. It is also possible to have other similarity cut offs, however it can be difficult to decide which threshold is the most suitable.

An attempt to compare the centroids and the cluster composition was performed using BLAST. The BLAST search output percent identity values for centroids that give a match in the paired sample comparisons, and centroids without a hit was assigned with 0% identity afterwards. The addition of the 0% identity was to account for differences in cluster numbers and composition. The search resulted in a boxplot (figure 3.3) showing the mean values for each paired comparison between methods. The results of the comparisons gave quite low identity values, but the most similar methods are Vsearch and Swarm with an average percent identity for the samples at around 60-70%. This result shows that Vsearch and Swarm presumably chose similar centroids, but also that their cluster numbers are not that different; as the cluster numbers figure (3.1) also show. Dada2 have a higher similarity to Dereplication than both Vsearch and Swarm, probably because Dada2 produce more clusters (as seen in figure 3.1). Because the centroids in Dereplication are the unique sequences all clusters originally are derived from, it might be expected that all methods would have high similarities with Dereplication. However, the addition of the non-hit clusters give a large amount of zeros in the computation of the mean because Dereplication have such large cluster numbers.

The 16S amplicons are in general very similar, hence without the consideration of the difference in cluster number, all comparisons would end up with a high percent identity. With this in mind the low average percent identity results of the BLAST search give an indication that method have big impact on the cluster composition. Consequently, method will probably also impact the core microbiota detection.

4.2 Clustering the total data set

In the hunt for a core microbiota it seemed more natural to cluster the entire data set in total, the clustering was performed as described in Methods. The resulting OTU tables ended up very different in terms of the number of OTUs. The difference in OTU number may say something about how stringent the clustering of the reads was. Vsearch is the least stringent and ended up with the smallest amount of OTUs (approx. 500), in other words Vsearch group more amplicons together in fewer clusters. Swarm have roughly double the number of clusters compared to Vsearch (approx. 1100), hence, Swarm is more stringent than Vsearch. The clustering with the Dada2 method was performed in a different manner than that of Vsearch and Swarm. The biggest difference is that Dada2 merges the forward and reverse reads after clustering them individually, as mentioned earlier. Another difference is that the samples are first clustered separately, and then all samples and OTUs are combined into an OTU table after clustering and merging. Dada2 ended up with approximately 7000 OTUs, and this result clearly show that Dada2 is clustering the amplicons with a much more stringent threshold than both Vsearch and Swarm. Dereplication is the most strict of the clustering methods used in this study, as it only clusters the unique sequences. The Dereplication method yielded nearly 700 000 OTUs. With this many OTUs downstream analyses will be very demanding, both in terms of cpu-time and computational demand.

Like the sample by sample clustering, cluster size (figure 3.4) was reported in addition to cluster number. Also in this result Vsearch and Swarm have a noticeable amount of very large clusters and a larger amount of smaller clusters. Dada2 on the other hand is more stable in size, and does not have any clear outliers. All methods vary quite a lot in size with both small and large clusters.

The larger clusters produced by both Vsearch and Swarm are more likely to be represented across a larger amount of samples, i.e. have high prevalence. The definition of core microbiota is that it has high prevalence in samples, hence a larger cluster size makes the cluster more probable of being part of the core. However, this is not always the case as there does not have to be a connection between large clusters and high prevalence. Prevalent clusters can have small abundances across a large amount of samples, and this would lead to a small cluster size and high prevalence. On the other hand, a large cluster can be highly abundant in few samples, thus having a large cluster size but low prevalence.

The rarefaction curves that were produced for all the methods (figure 3.5) display a kind of richness estimate for the different OTU tables. Also in these results does Vsearch and Swarm resemble each other, with a fast rising curve at first then moving over in the asymptotic phase. The asymptotic curve is a sign of reaching a state where no new OTUs are observed. Dada2 starts similar to the other two, but does not move from the fast-rising curve to the steady growing asymptotic curve. Instead, the curve transitions into a linear growing state. This behaviour can be a sign of noise in the data; new OTUs are constantly detected, regardless of the amount of additional new samples. One might speculate if the large number of clusters are what caused the Dada2 curve to look this particular way. However, the curve from the Dereplication table, with nearly 700 000 OTUs, behave in a more similar manner to Vsearch and Swarm with the curve transitioning over in the asymptotic state at the end. The Dada2 curve clearly stands out from the other three, which leads

me to assume that the Dada2 algorithm produce noisy results with too many false positives.

4.3 Core microbiota

In order to see if there was any signs of core microbiota for the four methods, the number of samples present in each cluster was counted (figure 3.6). Vsearch and Swarm were the only methods that had clusters with reads present in all samples. Dada2 reached a sample count just above 200 which is a prevalence below 50%. It was expected that Dada2 would have lower prevalence values than Vsearch and Swarm because of the increase in the number of clusters. A rather unexpected result is that Dereplication have clusters detected in over 300 samples, i.e. unique sequences present in more than 300 samples that Dada2 does not detect. The OTU number for Dereplication is massive compared to Dada2, hence the low prevalence values for Dada2 cannot be explained by a high cluster number.

A possible explanation can be that the sample by sample clustering procedure by Dada2 omit these highly prevalent sequences because they have a low abundance in each sample. Hence, they might be clustered with other sequences that are different in each sample and then when the samples are joined in the OTU table the centroids between the samples are not identical, i.e. not part of the same OTU. Even though the default setting to Dada2 clustering is sample by sample, because the sequence variants are resolved and a sequence variant should be constant across samples, it seems to be missing reads that appear in a large amount of samples. Thus, it might be an argument to pool the samples before clustering to detect these prevalent unique sequences that Dereplication recognize, and Dada2 overlook. At the cost of higher computational demand and time consumption.

4.3.1 Phylogenetic detection of the core

It was decided to create a phylogenetic tree, to further investigate which taxa were part of the core (figure 3.13). The nodes in the tree were colored according to prevalence. The tree was based on the Vsearch data, because the previous results displayed core OTUs. It was presumed that Vsearch and Swarm would produce similar trees based on the previous results, hence the Swarm tree was not produced. The Dada2 per cluster sample count showed that there were no core microbiota in the Dada2 OTU table. The tree incorporates the taxonomic information of the classification with the phylogenetic information between the centroids used to make the tree. The prevalence information is preserved in the coloring of the nodes. All taxonomic annotations used in the tree were at the genus level.

In the tree there are three genera that stand out as highly prevalent and possible members of the core microbiota across all samples; *Lactobacillus, Gilliamella* and *Snodgrassella*. These genera are proposed as part of the core microbiota in other studies as well (Kwong and Moran, 2016). The *Lactobacillus* genus are represented by many nodes in the prevalence tree, and all seem to be fairly closely related. Many of these nodes also have a quite high prevalence at around 75%. Hence *Lactobacillus* seems to be an important genus for the honey bee gut. However, there is only one *Lactobacillus* node that is colored in pink, indicating that there is one particular *Lactobacillus* species that is an important part of the core.

Nodes in the tree representing *Gilliamella, Snodgrassella* and *Frischella* seem to be closely related, and the nodes are intertwined indicating that the taxonomical classification struggle to separate these genera. All these three genera are taxonomically in the same phylum (*Proteobacteria*) and *Gilliamella* and *Frischella* are in the same family (*Orbaceae*), and this is likely the reason why the nodes are intertwined in the phylogenetic tree. For the *Gilliamella* and *Snodgrassella* genera, the same thing is observed as it was for the *Lactobacillus* genus. There are only one of the nodes from each that are colored in pink (have high prevalence), indicating that there is one species from each of these genera that are important in the core microbiota.

4.3.2 Finding core microbiota in the Dada2 data

Dada2 claims to be able to separate amplicons on the strain level (Callahan et al., 2016), and if this is correct it may be an explanation of why the method produce so many OTUs. There are studies that claim 16S rRNA is not able to separate bacteria well on a species level (Gevers et al., 2005; Fox et al., 1992). With the small pieces of the gene used in amplicon sequencing, identification of bacterial species is even harder. Because of this a decision was made to create a phylogenetic tree from the Dada2 OTU table were the read counts from OTUs belonging to each unique genus was merged, resulting in a table containing only the unique genera and their corresponding, summed up read counts (figure 3.14). In the resulting genus tree, none of the genera had a prevalence at 100%. The genera *Lactobacillus* and *Gilliamella* have a high prevalence (colored in pink), 93% and 95%, respectively. *Snodgrassella*, which were highly prevalent in the Vsearch tree (figure 3.13), had a prevalence of 86.9% in the Dada2 genus tree.

Analogous trees were created for the other three methods (Dereplication: figure 3.15; Vsearch/Swarm: figure A.1, A.2), and these are very different from the Dada2 genus tree. When the OTUs are merged to their genus-level in the other three OTU tables, a much higher portion of the nodes end up having a prevalence at 90% or higher compared to Dada2. If Dada2 were just compared to Vsearch and Swarm, the result would might not be that unexpected. However, when the Dereplication tree look more like the Vsearch and Swarm trees and Dada2 is this far from any of the other trees, it is very unexpected. The reads in the genera making up the pink nodes in the Vsearch/Swarm/Dereplication trees seem to be spread across many genera in the Dada2 table.

All OTU tables are created by the exact same input reads, and the centroids are classified using the same method. Dada2 produce more OTUs for all genera compared to Vsearch and Swarm (table 3.3). For example for the genus *Bifidobacterium*, Vsearch and Swarm infer 15 and 19 clusters, respectively. For the same genus, Dada2 infer 157 clusters. When merging the Vsearch and Swarm clusters (and Dereplication, with a count of 23 163 clusters) for this genus, it ends up being highly prevalent (have a pink node, $> \approx 90\%$ prevalence). When the 157 clusters from Dada2 is merged, the same thing does not happen. The genus end up with a prevalence of 70%, this is noticeably lower than the other three methods.

Combined with the results for the original Dada2 OTU table, the genus trees show that Dada2 have a poor ability to detect core OTUs and genera for this data set. Dada2 have considerably different results from the other methods regarding

core OTUs, and does not seem to be able to detect genera present in a large number of samples.

4.4 Effect of sampling categories

Honey bee guts can be divided into four parts; crop, midgut, ileum, and rectum. Hence, it was of interest to investigate if the different gut parts had differences in their microbiota composition. Especially for the Dada2 data, it was of interest to figure out if differences between gut parts were a factor causing the absence of core microbiota.

A starting point for the exploration of the gut parts was an MDS analysis based on the Vsearch OTU table. Weighted UniFrac were used as distances in the analysis to see if there were any grouping of the samples based on any of the sampling categories. Figure 3.7 displays the result of the analysis, and there is clearly no grouping based on which hive the samples were taken from. Therefore hive was not included as a grouping factor in further analyses. The MDS figure show that there are signs of grouping based on gut part, however the separation between groups are not perfectly clear.

In addition to the MSD analysis, alpha diversity were calculated for each month in each gut part for all methods (figure 3.8). The diversity measure is rather stable, for each method, across months in the different gut parts, and there is not very large variation in mean between gut parts either. Vsearch and Swarm have nearly identical diversity values across all gut parts, even though Swarm have double the amount of OTUs compared to Vsearch. The Shannon entropy measure used to calculate the alpha diversity considers the number of OTUs in the calculation. Hence, it is not surprising that both Dada2 and Dereplication have higher mean values. Dada2 and Dereplication agree on most of the changes in diversity, and also change more than Vsearch and Swarm. Something to note is that in the crop fall samples (August and November), Dada2 have a decrease in diversity while Vsearch and Swarm show an increase. If the methods were used individually this difference would lead to completely opposite conclusions regarding the change in diversity during the fall. Vsearch and Swarm diversities vary a lot less than those of Dada2, there is much more stability in the diversity estimations across months. Dada2 values have a large standard deviation, especially in the later months, indicating variation in the diversity measures between samples in the same month and gut part. Overall, it does not seem to be a noticeable difference in diversity between gut parts, only slight differences between months within a gut part. However, the monthly differences and changes detected are influenced by the clustering methods.

Further, as for the total data, the number of samples present in each cluster was counted, however the OTU table was divided based on gut part beforehand (figure 3.9, 3.10, 3.11 & 3.12). For Vsearch and Swarm there are more core OTUs than for the total data set, as one might expect when there are fewer possible samples for each gut part than for the whole data set. There are OTUs with 100% prevalence for all gut parts for both the methods. Dada2 and Dereplication does not have any OTUs present in all samples, and Dada2 is not anywhere near of having OTUs with 100% prevalence for any of the gut parts. The difference in OTU composition between gut parts does not seem to be the reason for the absence of core OTUs in the Dada2 table. It is as expected that Dereplication does not have any core

microbiota, because of the large amount of OTUs and the nature of the method. Also for the gut part per cluster sample count figures, Dereplication have clusters that are detected in more samples than in Dada2.

Regarding the phylogenetic analysis, the result of the gut part trees were not that different from the total Vsearch tree (figure 3.13). In ileum and rectum the same genera with high prevalence are detected as in the total tree (table 3.2). For midgut and crop the result is slightly different, and some other genera are identified as highly prevalent. In the midgut the same three genera as the total data are found, in addition to *Pseudomonas* and *Parasaccharibacter*. Crop is the gut part with the biggest difference from the total tree, with only *Gilliamella* as shared result. The additional genera found as highly prevalent in crop are; *Pseudomonas*, *Sphingomonas*, *Pseudoalteromonas* and *Parasaccharibacter*.

4.5 Future research

Because the results in this thesis reveal such large variation between methods for this data set, it would be wise to perform comparisons using simulated data sets. The mock communities should vary in size, composition and sequence quality to detect if the methods introduce any bias to particular types of data. Other biological data sets should also be tested for the same effects shown in this data set, to see if the results obtained in this thesis are specific for this particular data. If possible it would also be of interest to track each unique read through the clustering and understand where the differences in the grouping of reads occur. And to examine similarities in clustering without just being dependent on the taxonomical classification. Dada2 should also be run with the pooling option to inspect if the detection of core OTUs improve when clustering all samples at once.

Chapter 5

Conclusion

In this study the aim was to understand the differences between a set of amplicon clustering methods, and the effect of clustering in the detection of core microbiota in the honey bee gut. There are substantial differences between methods, which shows that the clustering of amplicon sequences is highly dependent on the method. The results from any study will always be affected by the method in some degree, and considerations must be made depending on the wanted output from an analysis.

The data shown in this study illustrate the importance of the method when studying specific topics like the core microbiota. Both Vsearch and Swarm were able to detect core OTUs in the total and divided by gut part data. Based on previous studies, the honey bee core detected by Vsearch and Swarm were as expected. Dada2 were not able to detect a core, and the method display a poor ability to distinguish microbiota present in a high percentage of samples. Overall, Dada2 seems to partition the data in too small units and, maybe, exaggerating the amount of OTUs present in the data. It might perform differently for different kinds of data, and these observations and conclusions are limited to the data set used in this study. Even though Dereplication, surprisingly, detected OTUs present in more samples than Dada2, the vast amount of OTUs are very unpractical to work with and the analyses on these data are very demanding in time consumption and computational space.

Lastly there were not detected any large effects of sampling categories; hive, month or gut part. From this study the core microbiota does not seem to have large variation depending on which time of the year it is, or which gut part it is sampled from. Crop were the only gut part standing out from the other three, that had a noticeable difference in core genera compared to the other gut parts.

Bibliography

D. Aguirre de Cárcer. The human gut pan-microbiome presents a compositional core formed by discrete phylogenetic units. *Sci. Rep.*, 8(1):14069, dec 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-32221-8.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, oct 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

K. E. Anderson, T. H. Sheehan, B. J. Eckholm, B. M. Mott, and G. DeGrandi-Hoffman. An emerging paradigm of colony health: microbial balance of the honey bee and hive (Apis mellifera). *Insectes Soc.*, 58(4):431–444, nov 2011. doi: 10.1007/s00040-011-0194-6.

A. Behnke, M. Engel, R. Christen, M. Nebel, R. R. Klein, and T. Stoeck. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ. Microbiol.*, 13(2):340–349, feb 2011. doi: 10.1111/j.1462-2920.2010.02332.x.

E. Boon, C. J. Meehan, C. Whidden, D. H.-J. Wong, M. G. Langille, and R. G. Beiko. Interactions in the microbiome: Communities of organisms and communities of genes. *FEMS Microbiology Reviews*, 38(1):90–118, jan 2014. ISSN 01686445. doi: 10.1111/1574-6976.12035.

J. Bunge and M. Fitzpatrick. Estimating the Number of Species: A Review. *J. Am. Stat. Assoc.*, 88(421):364–373, mar 1993. doi: 10.1080/01621459.1993.10594330.

B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, jul 2016. ISSN 15487105. doi: 10.1038/nmeth.3869.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009. doi: 10.1186/1471-2105-10-421.

J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335, apr 2010. doi: https://doi.org/10.1038/nmeth.f.303.

S. Chakravorty, D. Helb, M. Burday, N. Connell, and D. Alland. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods*, 69(2):330–339, may 2007. ISSN 0167-7012. doi: 10.1016/J.MIMET.2007.02.005.

J. Chen. *GUniFrac: Generalized UniFrac Distances*, 2018. URL `https://CRAN.R-project.org/package=GUniFrac`. R package version 1.1.

J. Chen, K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman, and H. Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28 (16):2106–2113, aug 2012. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts342.

F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, oct 2004. ISSN 00280836. doi: 10.1038/nature03001.

A. Criscuolo and O. Gascuel. Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC Bioinformatics*, 9(1):166, dec 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-166.

R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, mar 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh340.

R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, aug 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq461.

R. C. Edgar. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, page 081257, 2016. doi: 10.1101/081257.

R. C. Edgar and H. Flyvbjerg. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31(21):3476–3482, nov 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv401.

R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, aug 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr381.

A. Escobar-Zepeda, A. Vera-Ponce de León, and A. Sanchez-Flores. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front. Genet.*, 6:348, dec 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00348.

F. Finotello, E. Mastrorilli, and B. Di Camillo. Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Brief. Bioinform.*, 19(4): 679–692, dec 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw119.

G. E. Fox, J. D. Wisotzkey, and P. J. Jurtshuk. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.*, 42(1):166–170, jan 1992. doi: 10.1099/00207713-42-1-166.

E. Genersch. Honey bee pathology: current threats to honey bees and beekeeping. *Appl. Microbiol. Biotechnol.*, 87(1):87–97, jun 2010. ISSN 0175-7598. doi: 10.1007/s00253-010-2573-8.

D. Gevers, F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. V. de Peer, P. Vandamme, F. L. Thompson, and J. Swings. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.*, 3(9):733–739, 2005. doi: 10.1038/nrmicro1236.

N. J. Gotelli and R. K. Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4(4):379–391, jul 2001. doi: 10.1046/j.1461-0248.2001.00230.x.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer, New York, NY, 2nd edition, 2009. ISBN 978-0-387-84858-7. doi: https://doi.org/10.1007/978-0-387-84858-7. URL `https://web.stanford.edu/ hastie/Papers/ESLII.pdf`.

A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U. S. A.*, 111(13):4904–9, apr 2014. ISSN 1091-6490. doi: 10.1073/pnas.1402564111.

L. W. Hugerth and A. F. Andersson. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Frontiers in Microbiology*, 8:1561, sep 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.01561.

A. Khosravi and S. K. Mazmanian. Disruption of the gut microbiome as a risk factor for microbial infections. *Current Opinion in Microbiology*, 16(2):221–227, apr 2013. ISSN 1369-5274. doi: 10.1016/J.MIB.2013.03.009.

R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.*, 16(7): 410–422, jul 2018. ISSN 1740-1526. doi: 10.1038/s41579-018-0029-9.

A. F. Koeppel and M. Wu. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.*, 41(10): 5175–5188, may 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt241.

A. Konopka. What is microbial community ecology? *The ISME Journal*, 3(11): 1223–1230, nov 2009. ISSN 1751-7362. doi: 10.1038/ismej.2009.88.

W. K. Kwong and N. A. Moran. Gut microbial communities of social bees, jun 2016. ISSN 17401534.

N. Larsen, F. K. Vogensen, F. W. J. van den Berg, D. S. Nielsen, A. S. Andreasen, B. K. Pedersen, W. A. Al-Soud, S. J. Sørensen, L. H. Hansen, and M. Jakobsen. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic

Adults. *PLoS One*, 5(2):e9085, feb 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009085.

H. R. Lindman. *Analysis of Variance in Experimental Design*. Springer-Verlag New York, 1 edition, 1992. ISBN 0-387-97571-3. doi: 10.1007/978-1-4613-9722-9.

C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–35, dec 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.12.8228-8235.2005.

C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73(5):1576–85, mar 2007. ISSN 0099-2240. doi: 10.1128/AEM.01996-06.

T. Madden. The BLAST sequence analysis tool. *NCBI Handbook [Internet]. 2nd Edition.*, mar 2013. URL `https://www.ncbi.nlm.nih.gov/books/NBK153387/`.

F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, sep 2014. ISSN 2167-8359. doi: 10.7717/peerj.593.

F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420, 2015. doi: 10.7717/peerj.1420.

C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca, and J. Dore. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–11, feb 2006. ISSN 0017-5749. doi: 10.1136/gut.2005.073817.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, sep 2005. ISSN 0028-0836. doi: 10.1038/nature03959.

V. G. Martinson, J. Moy, and N. A. Moran. Establishment of characteristic gut bacteria during development of the honeybee worker. *Appl. Environ. Microbiol.*, 78(8):2830–40, apr 2012. ISSN 1098-5336. doi: 10.1128/AEM.07810-11.

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. *vegan: Community Ecology Package*, 2019. URL `https://CRAN.R-project.org/package=vegan`. R package version 2.5-5.

A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and biology insights*, 9:75–88, 2015. ISSN 1177-9322. doi: 10.4137/BBI.S12462.

E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2018. R package version 5.3.

J. E. Powell, V. G. Martinson, K. Urban-Mead, and N. A. Moran. Routes of acquisition of the gut microbiota of the honey bee Apis mellifera. *Appl. Environ. Microbiol.*, 80(23):7378–7387, dec 2014. ISSN 10985336. doi: 10.1128/AEM.01861-14.

C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(12):833–844, 2017. ISSN 15461696. doi: 10.1038/nbt1217-1211b.

T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, oct 2016. ISSN 2167-8359. doi: 10.7717/peerj.2584.

M. J. Rosen, B. J. Callahan, D. S. Fisher, and S. P. Holmes. Denoising PCR-amplified metagenome data. *BMC bioinformatics*, 13:283, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-283.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040454.

F. Sánchez-Bayo and K. A. Wyckhuys. Worldwide decline of the entomofauna: A review of its drivers. *Biol. Conserv.*, 232:8–27, apr 2019. ISSN 0006-3207. doi: 10.1016/J.BIOCON.2019.01.020.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–7, dec 1977. ISSN 0027-8424. doi: 10.1073/pnas.74.12.5463.

M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, 43(6):e37–e37, mar 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1341.

P. D. Schloss and J. Handelsman. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology*, 6(8):229, aug 2005. ISSN 14656906. doi: 10.1186/gb-2005-6-8-229.

C. E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

L. Snipen and K. H. Liland. *microseq: Basic Biological Sequence Handling*, 2018. URL `https://CRAN.R-project.org/package=microseq`. R package version 1.2.2.

F. Sommer and F. Bäckhed. The gut microbiota-masters of host development and physiology. *Nature Reviews Microbiology*, 11(4):227–238, 2013. ISSN 17401526. doi: 10.1038/nrmicro2974.

J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, 5(6):729–731, 1988. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040527.

K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3):512–526, may 1993. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a040023.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, 2007. ISSN 14764687. doi: 10.1038/nature06244.

Y. Van de Peer, S. Chapelle, and R. De Wachter. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.*, 24(17):3381–3391, sep 1996. ISSN 13624962. doi: 10.1093/nar/24.17.3381.

E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, sep 2014. ISSN 0168-9525. doi: 10.1016/J.TIG.2014.07.001.

T. Větrovský and P. Baldrian. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, 8(2):e57923, feb 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0057923.

G. M. Weinstock. Genomic approaches to studying the human microbiota. *Nature*, 489(7415):250–256, sep 2012. ISSN 0028-0836. doi: 10.1038/nature11553.

C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.*, 74(11):5088–90, nov 1977. ISSN 0027-8424. doi: 10.1073/pnas.74.11.5088.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, mar 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-3-r46.

Y. Yu, C. Lee, J. Kim, and S. Hwang. Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol. Bioeng.*, 89(6):670–679, mar 2005. ISSN 0006-3592. doi: 10.1002/bit.20347.

Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, D. Wan, R. Jiang, L. Su, Q. Feng, Z. Jie, T. Guo, Z. Xia, C. Liu, J. Yu, Y. Lin, S. Tang, G. Huo, X. Xu, Y. Hou, X. Liu, J. Wang, H. Yang, K. Kristiansen, J. Li, H. Jia, and L. Xiao. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.*, 37(2):179–185, feb 2019. ISSN 1087-0156. doi: 10.1038/s41587-018-0008-8.

Appendices

Appendix A

Phylogenetic trees

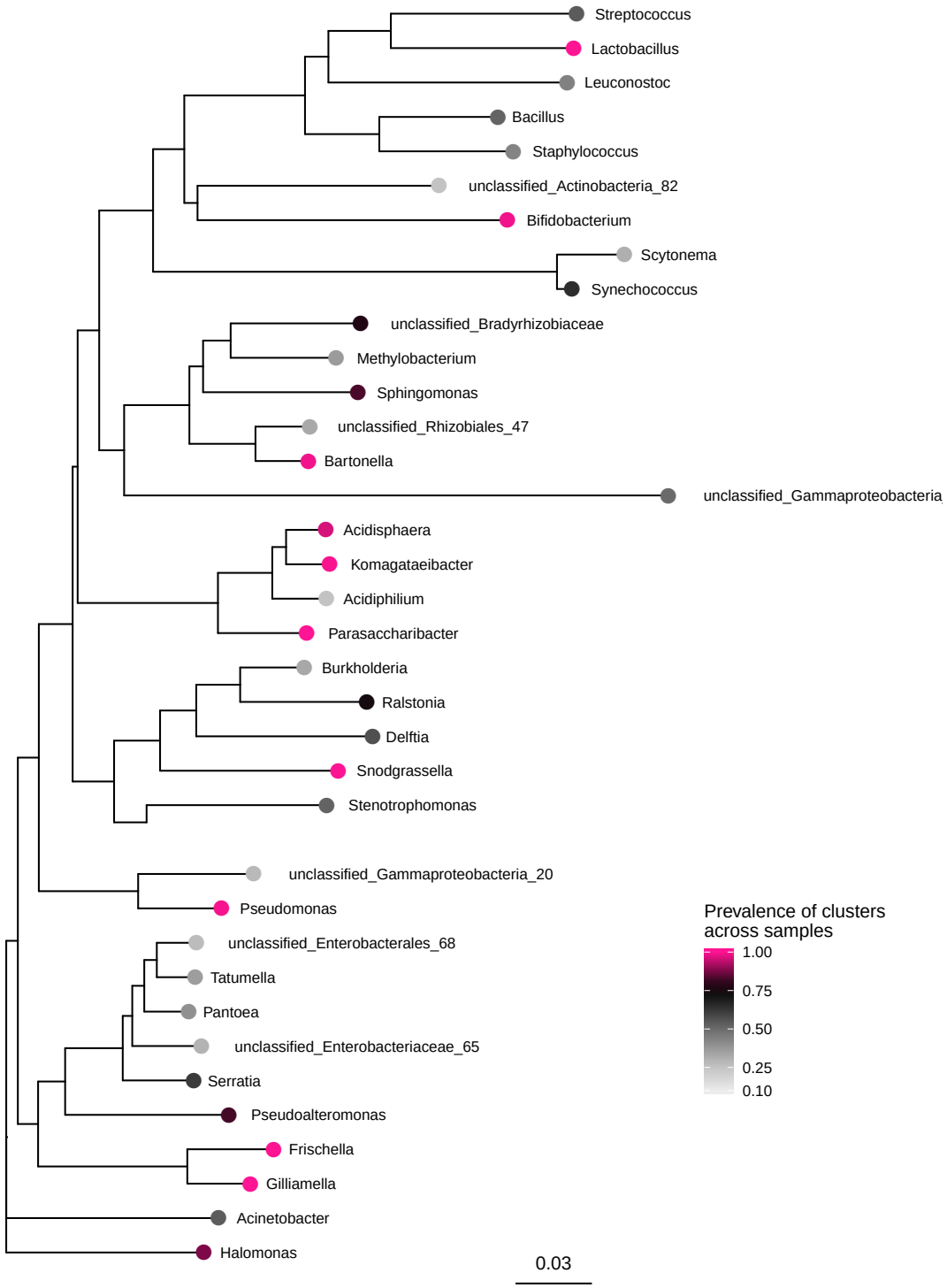


Figure A.1: Merged genera prevalence tree from the Vsearch OTU table. The colored tip represents the prevalence of the clusters across all samples. To reduce the number of clusters in the tree the OTUs with a prevalence below 25% are not included. The leaves are labeled by genera.

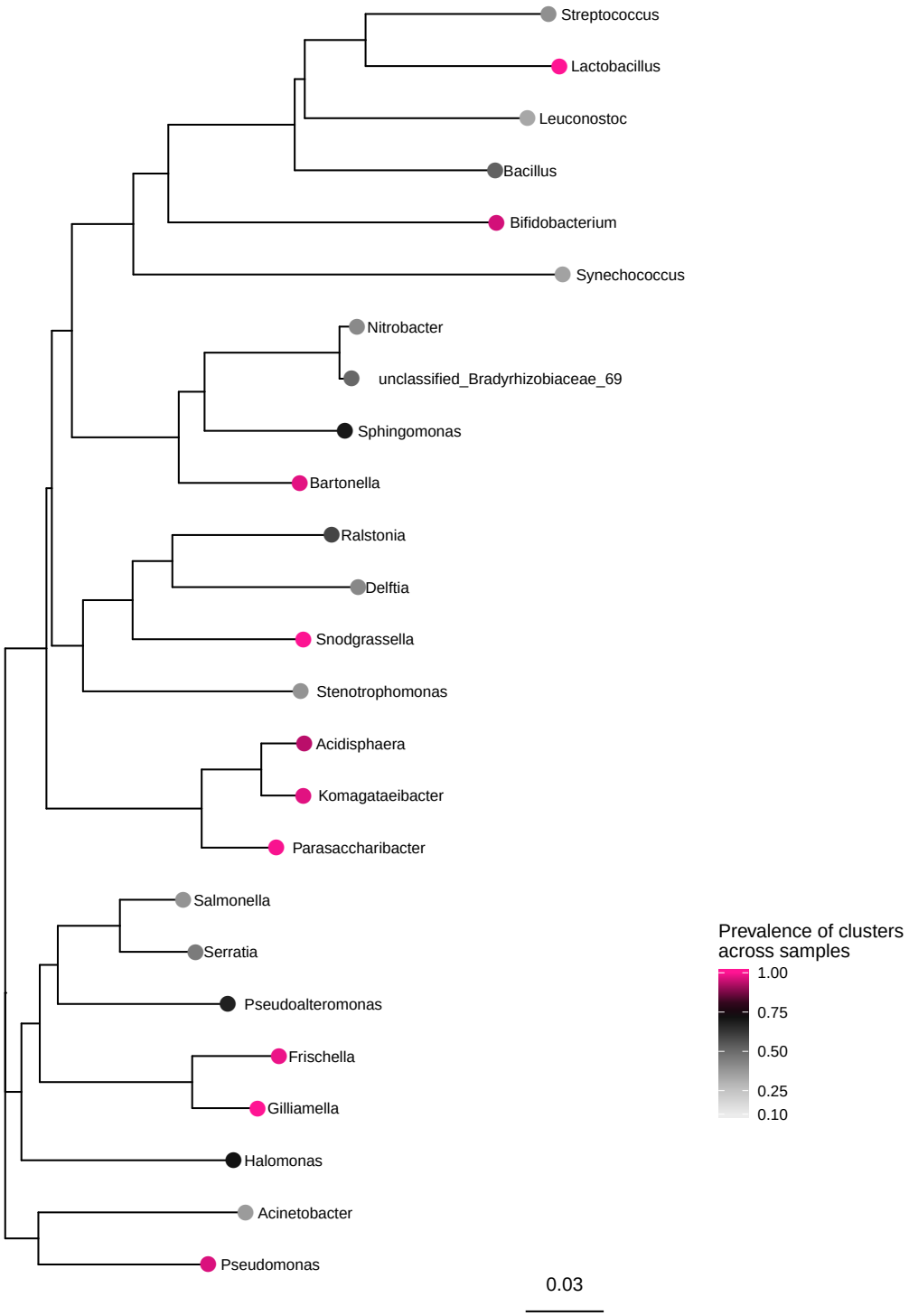


Figure A.2: Merged genera prevalence tree from the Swarm OTU table. The colored tip represents the prevalence of the clusters across all samples. To reduce the number of clusters in the tree the OTUs with a prevalence below 25% are not included. The leaves are labeled by genera.

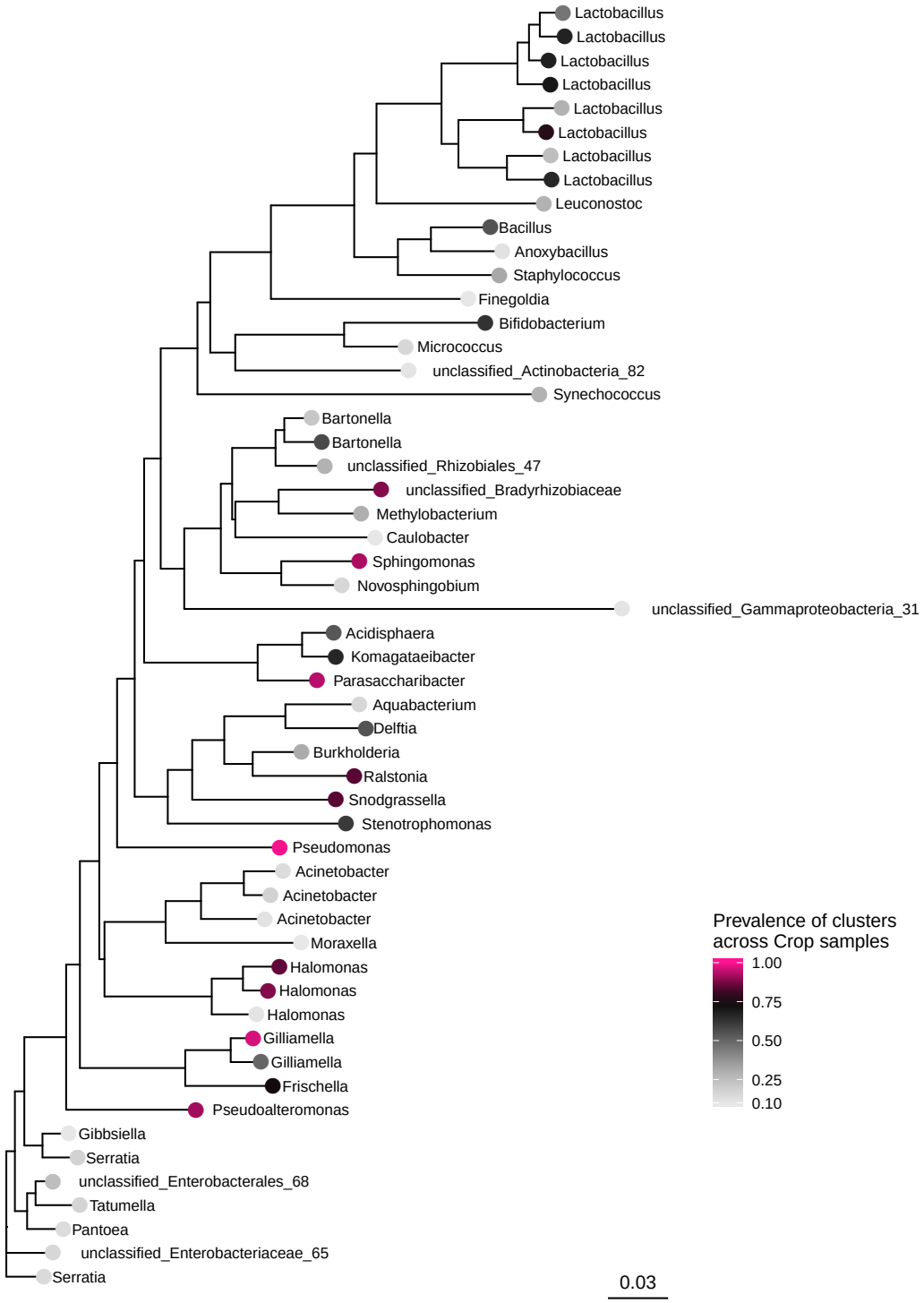


Figure A.3: Prevalence tree for the Crop samples from the Vsearch OTU table. The colored tip represents the prevalence of the cluster across all samples, the number of Crop samples is 100. Only the top 99% of the reads for each sample are included. To reduce the number of clusters in the tree the OTUs with a prevalence below 10% are not included. The leaves are labeled by genera.

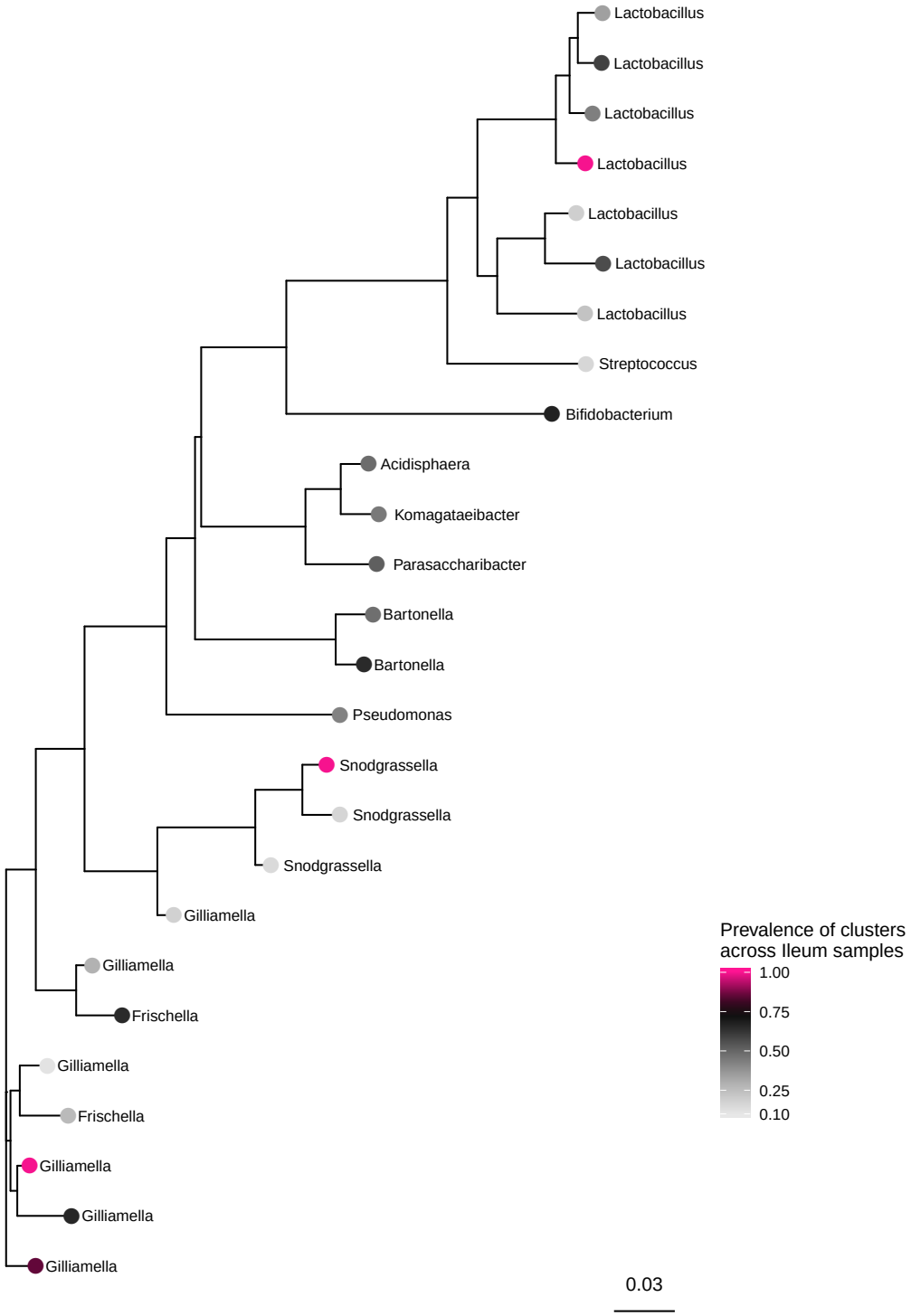


Figure A.4: Prevalence tree for the Ileum samples from the Vsearch OTU table. The colored tip represents the prevalence of the cluster across all samples. There are 120 Ileum samples. Only the top 99% of the reads for each sample are included. To reduce the number of clusters in the tree the OTUs with a prevalence below 10% are not included. The leaves are labeled by genera.

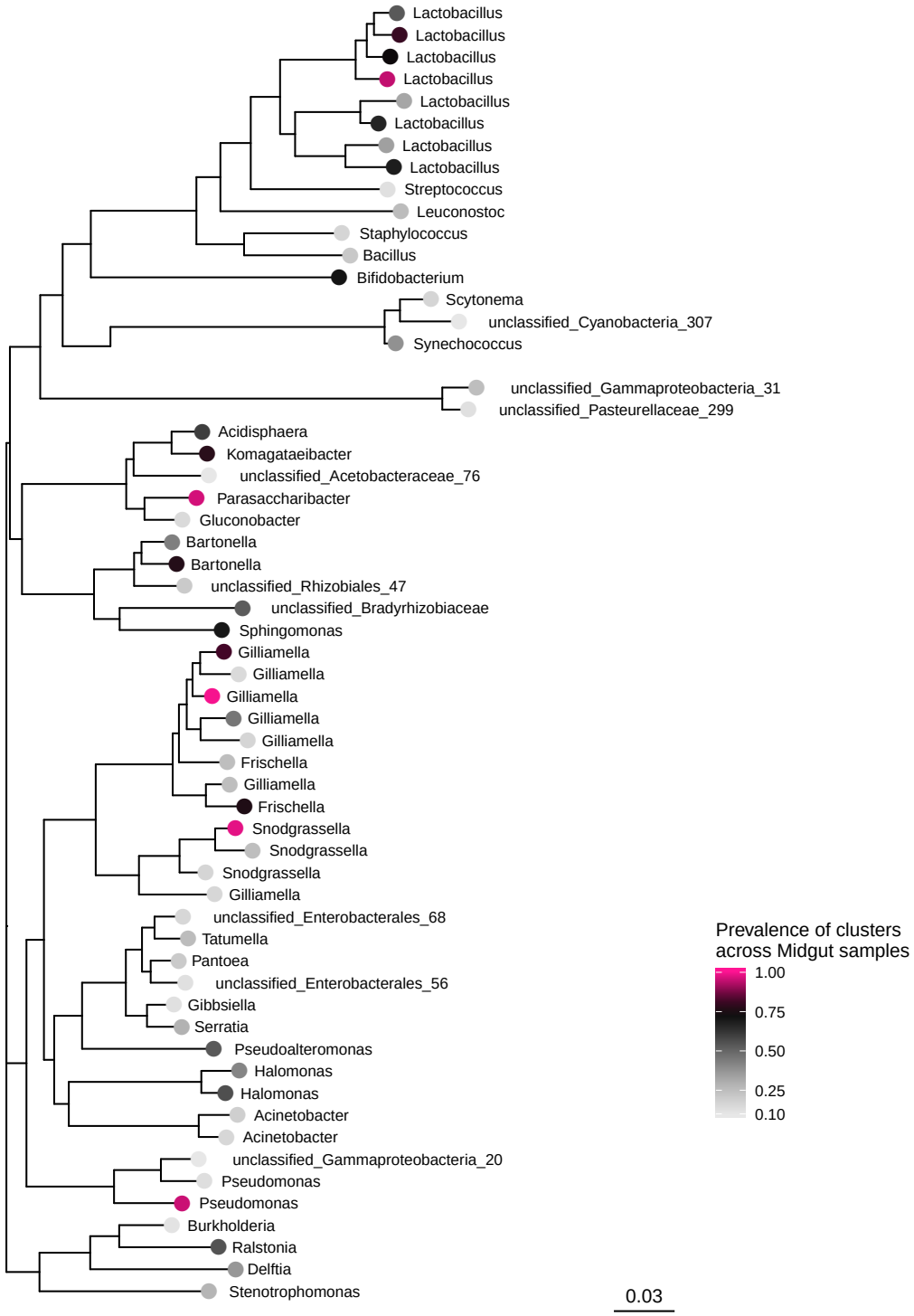


Figure A.5: Prevalence tree for the Midgut samples from the Vsearch OTU table. The number of samples from the Midgut is 120. The colored tip represents the prevalence of the cluster across all samples. Only the top 99% of the reads for each sample are included. To reduce the number of clusters in the tree the OTUs with a prevalence below 10% are not included. The leaves are labeled by genera.

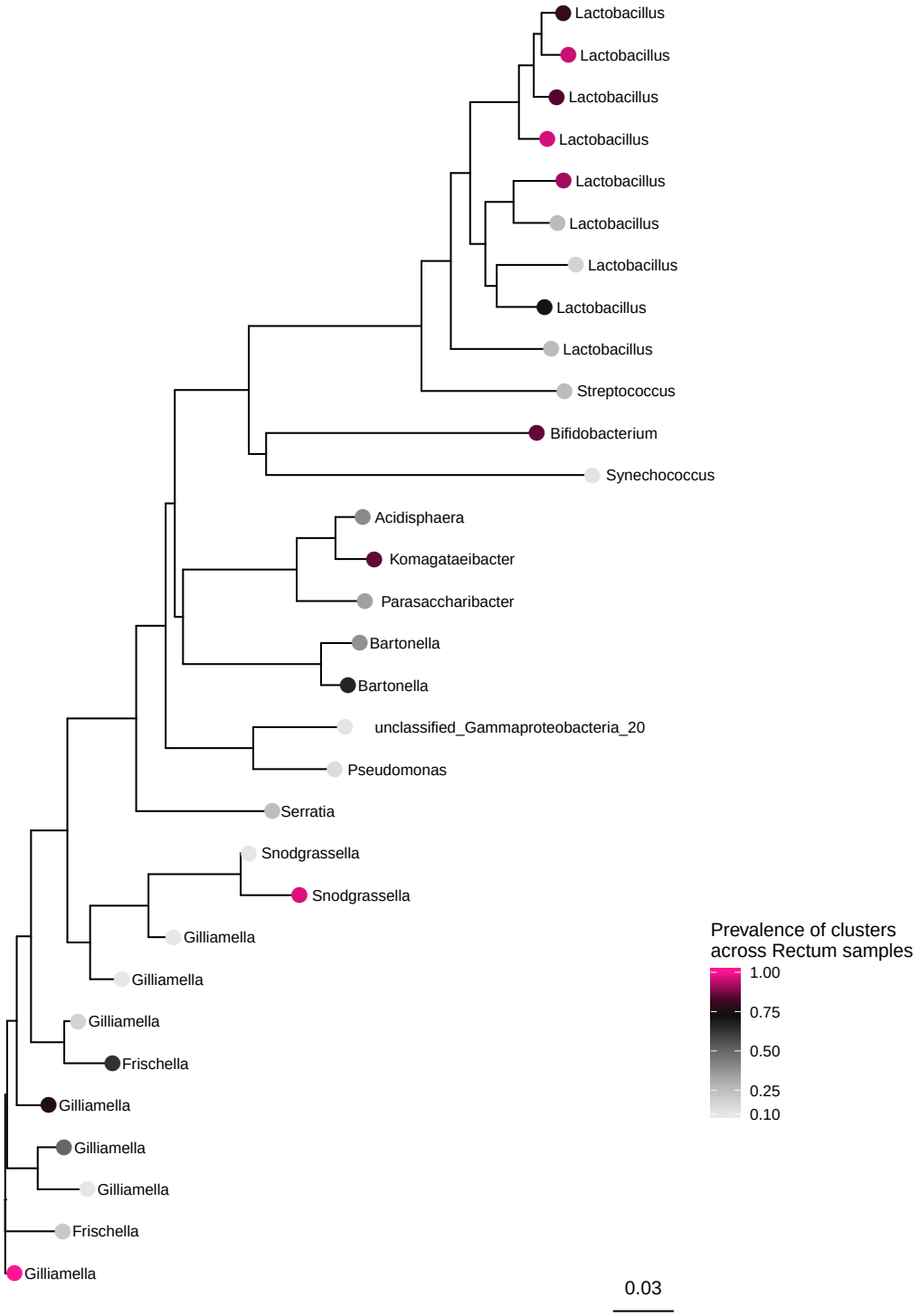
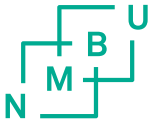


Figure A.6: Prevalence tree for the 120 Rectum samples from the Vsearch OTU table. The colored tip represents the prevalence of the cluster across all samples. Only the top 99% of the reads for each sample are included. To reduce the number of clusters in the tree the OTUs with a prevalence below 10% are not included. The leaves are labeled by genera.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway