DISSERTATIONES
TECHNOLOGIAE
CIRCUMIECTORIUM
UNIVERSITATIS
TARTUENSIS
27

JENS-KONRAD PREEM

Forest soil bacterial community analysis using high-throughput amplicon sequencing

DISSERTATIONES TECHNOLOGIAE CIRCUMIECTORIUM
UNIVERSITAS TARTUENSIS
27

DISSERTATIONES TECHNOLOGIAE CIRCUMIECTORIUM
UNIVERSITAS TARTUENSIS
27

JENS-KONRAD PREEM

Forest soil bacterial community analysis using high-throughput amplicon sequencing

UNIVERSITY OF TARTU
Press

Department of Geography, Institute of Ecology and Earth Sciences, Faculty of Science and Technology, University of Tartu, Estonian

This dissertation has been accepted for commencement of the degree of Doctor of Philosophy in Environmental Technology on September 11, 2017 by the Scientific Council on Environmental Technology of the Faculty of Science and Technology of the University of Tartu

Supervisors: Prof. Dr. Jaak Truu, Department of Geography, University of Tartu, Estonia

Prof. Dr. Ülo Mander, Department of Geography, University of Tartu, Estonia

Opponent: Timo Petteri Sipilä, PhD, Institute for Molecular Medicine Finland

Commencement: November 10th, 2017, at 14:15 in room 327 (J.G. Granö auditorium), University of Tartu, 46 Vanemuise Street, Tartu

The publication of this dissertation has been funded by the Department of Geography of the University of Tartu.

ISSN 1736-3349
ISBN 978-9949-77-576-7 (print)
ISBN 978-9949-77-577-4 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

ORIGINAL PUBLICATIONS

I. **Preem J.-K.**, Truu J., Truu M., Mander Ü., Oopkaup K., Lõhmus K., Helmisaari H.-S., Uri V., Zobel, M., 2012. Bacterial community structure and its relationship to soil physico-chemical characteristics in alder stands with different management histories. Ecological Engineering, 49, 10–17. http://doi.org/10.1016/j.ecoleng.2012.08.034

II. Truu M., Ostonen I., **Preem J.-K.,** Lõhmus K., Nõlvak H., Ligi T., Rosenvald K., Parts K., Kupper P., Truu J., 2017. Elevated air humidity changes soil bacterial community structure in the silver birch stand. Frontiers in Microbiology, 8, 1–15. http://doi.org/10.3389/fmicb.2017.00557

III. Ostonen I., Truu M.,Helmisaari H.-S., Lukac M., Borken W., Vanguelova E., Godbold D. L., Lõhmus K., Zang U., Tedersoo L., **Preem J.-K.**, Rosenvald K., Aosaar J., Armolaitis K., Frey J, Kabral N., Kukumägi M., Leppälammi-Kujansuu J., Napa Ü., Nöjd P., Lindroos A.-J., Merilä P., Parts K., Uri V., Varik M. Truu J., 2017. Adaptive root foraging strategies along a boreal – temperate forest gradient, New Phytologist 215(3): 977–991 http://doi.org/10.1111/nph.14643

Author's contribution

Publication I: The author was responsible for sequence data processing (100%), analysis (80%) and writing of the manuscript (60%)

Publication II: The author was responsible for sequence data processing (100%) and data analysis, including statistical analysis (80%), and writing the manuscript (20%)

Publication III: The author was responsible for 16S sequence data processing (100%) and 16S sequence data analysis including statistical analysis (80%), and writing the manuscript (20%)

1. INTRODUCTION

The soil as a central agent in many ecological processes has received a lot of research attention from many different angles. The investigation of the rich microbiome of the soil has been slowed by the fact that most of the microbes are unculturable. This gap can be filled by the metagenomics which is a field that deals with genetic material directly acquired form environmental samples offering a new way to investigate soil microbiome.

High-throughput sequencing of PCR-amplified 16S rRNA gene has grown explosively over the past decade. The analysis of 16S rDNA data usually begins with the construction of operational taxonomic units (OTUs): clusters of reads that differ by less than a fixed sequence dissimilarity threshold. Consequently, the obtained sample-by-OTU abundance table serves as the basis for further statistical and exploratory analysis. For example, 16S rDNA sequences clustered into OTU-s could be used to estimate the microbial community richness, alpha and beeta diversity, and composition of bacterial communities or be used in different statistical and multivariate exploratory analyses as a stand in for more conventional taxonomic units.

The rapid development of sequencing technologies and amount of 16S rDNA data has been coherent with the fast growth in different analytical tools and programs for analysis of obtained data sets. During the last decade, a plethora of tools based on different principles and having different computational requirements to perform aforementioned OTU clustering has been created (Schloss, 2016).

There have been some comparison and benchmarking studies for these tools and methods (Sun et al., 2012; Chen et al., 2013). In this work we are not so much interested in direct benchmarking but in the differences of the final outcome of series of analyses when different OTU clustering methods are used.

The original publications on which this work is based concern with forest soil microbiome and the associated plant microbiomes (specifically rhizosphere microbiome). Methods that were used in these publications were based on 16S rDNA amplicon based metagenomics techniques.

2. THE AIM OF THE STUDY

In this work we used the dataset published in Preem et al. (2012; Publication I) and analysed it using different software packages for processing bioinformatics data: Mothur (Schloss et al. 2009), UCLUST (Edgar 2010), CROP (Hao et al. 2011), and a novel Swarm (Mahé et al. 2014). The results were compared with the datasets on microbiological and environmental parameters. The main aim of the study was to evaluate the different clustering methods for their use in 16S rDNA metagenomic analyses and observe the effect of different OTU binning methods on the final output and ecological conclusions of metagenomic analysis.

In addition, to better evaluate the differences between different clustering methods in silico mock community analysis was also performed. An OTU clustering independent analysis – a Principal Coordinates Analysis (PCoA) based on Kantorovich-Rubenstein distances and acquired from phylogenetic placements – is also provided as a background information. In this analysis, sequence data from other original publications were used.

3. LITERATURE REVIEW

3.1. Forest soil microbiome

Microbial communities are vital in mediating the forest soil biogeochemical cycles, and an understanding of their role in ecosystem processes is pivotal predicting the forest response to future environmental conditions. Fungi are the most well-studied microbes in temperate and boreal forests' soils, while bacteria and archaea represent another important but less explored integral part of the of the forest soil microbial community (Baldrian 2016).

Bacteria inhibit several habitats in forest soils – they can be found in bulk soil, rhizosphere, and litter. The collective communities of plant-associated microorganisms are referred to as the plant microbiome. Virtually all tissues of a plant host a microbial community that is a part (endosphere) of the plant microbiome (Turner et al. 2013), and rhizosphere microbes constitute another part of the plant microbiome (Mendes et al. 2013). Despite its complexity and dynamism, particularly in natural environments, it is important not to overlook the plant microbiome when performing studies about forest soil microbial communities.

Typical dominant bacterial phyla in forest soils are *Acidobacteria, Actinobacteria, Proteobacteria, Bacteroidetes*, and *Firmicutes* (Lladó et al. 2017). Archaeal community is mostly dominated by phylum *Thaumarchaeota*, while *Euryarchaeota* and *Crenarchaeota* are less abundant (Siles and Margesin 2016).

Forest soils as a habitat for bacteria is very heterogeneous as these soils are characterized by sharp vertical stratification. Along the vertical soil gradient the quantity and quality of organic matter is decreasing, and this affects the soil bacterial and archaeal community structure, abundance and activity (Norman and Barrett, 2016). In addition to soil organic matter content and quality the soil pH value is important factor shaping forest soil bacterial community structure (Jeanbille et al. 2016; Nacke et al. 2016). Furthermore, soil parameters like availability of organic and inorganic nitrogen and phosphorous is related to forest soil microbial community structure (Chodak et al. 2015) Also, the soil type may have impact on soil microbial community structure (Colin et al. 2017).

Trees strongly shape the community composition of soil bacteria and fungi in temperate and boreal forests (Nacke et al. 2016; Uroz et al. 2016a). Uroz and co-workers (2016b) found that fungal and archaeal community structures and compositions were mainly determined according to tree species, whereas bacterial communities differ to a great degree between rhizosphere and bulk soils, regardless of the tree species in beech and Norway spruce stands.

Microbial communities in forest soils respond to the effects of global change, such as climate warming, increased levels of carbon dioxide, or anthropogenic nitrogen deposition (Lladó et al. 2017). Soil microbial communities are able to respond more rapidly than plant communities to environmental changes,

which in turn affect ecosystem processes, such as carbon and nitrogen cycling, because of the vastness of microbial biomass and diversity (López-Lozano et al. 2013). Since soil microbes (primarily prokaryotes and fungi) produce greenhouse gases, especially N_2O (Seo and DeLaune 2010; Giles et al. 2012; Saggar et al. 2013) and methane (Nazaries et al. 2013a) climate change-triggered alterations in soil microbial communities can have substantial feedback to the climate (Nazaries et al. 2013b). Climate warming is increasingly leading to marked changes in plant and animal biodiversity, but it remains unclear how temperatures affect microbial biodiversity, particularly in terrestrial soils (Zhou et al. 2016).

In accordance with metabolic theory of ecology, taxonomic and phylogenetic diversity of soil bacteria, fungi and nitrogen fixers are all better predicted by variation in environmental temperature than pH. Cong and co-workers (2015) found that temperature was important in shaping microbial communities at both the taxonomic and functional gene levels in different forest soils. However, the rates of diversity turnover across the global temperature gradients are substantially lower than those recorded for trees and animals, suggesting that the diversity of plant, animal and soil microbial communities show differential responses to climate change. (Zhou et al. 2016).

3.2. Soil microbiome characterization using metagenomics methods

3.2.1. General information on metagenomics

Metagenomics refers to the use of genomic techniques on the genetic material acquired directly from environment – without intermediate isolation and cultivation of individual species. To our knowledge the earliest publication using the term metagenome is by Jo Handelsman (Handelsman et al. 1998). It is remarkable that the idea of metagenomics (the term is in this publication referencing the idea of analysing collection of genes sequenced from the environment in a way analogous to the study of a single genome) is first mentioned in the context of soil microbes.

The same article (Handelsman et al. 1998) brings out a combination of factors that make it very compelling to use metagenomic methods on soils – first: a great proportion of soil microbes are unculturable, second: the soil microbial communities have shown to be extremely rich and diverse containing not only a vast amount of different taxa but also a vast amount of different metabolic pathways and functionalities. The latter making it an enticing proposition for practical/applied research.

The main questions to answer in the microbial ecology are “Who is out there?” and “What are they doing?” In fact, metagenomics can answer both questions. Particularly, microbial diversity can be determined using two different approaches: (1) Amplicon sequencing or (2) Shotgun metagenomics. In the

first approach, specific regions of DNA from microbial communities are amplified using taxonomical informative primer targets such as 16S rRNA gene for prokaryotes and intergenic transcribed spacers (ITS) or the large ribosomal subunit (LSU) gene for eukaryotes (Sharpton 2014; Tonge et al. 2014). In the second approach, shotgun metagenomics can help to reconstruct large fragments or even complete genomes from microorganisms in a community without previous isolation, allowing the characterization of a large number of coding and non-coding sequences that can be used as phylogenetic markers (Escobar-Zepeda et al. 2015).

Both approaches have their strengths and weaknesses and it must be kept in mind that the end results using one or other method might disagree on with each other as has been shown in study by Steven and co-workers (Steven et al. 2012).

3.2.2. Amplicon based metagenomics/metaprofiling

As previously mentioned amplicon based approach uses specific regions of DNA from communities that are amplified using taxonomical informative primer targets, and in case of prokaryotes 16S rRNA gene is the most common example. All original publications (Publication I (Preem et al. 2012), Publication II (Truu et al. 2017), and Publication III (Ostonen et al. 2017)) serving as a basis of this doctoral thesis include 16S rRNA amplicon based metagenomics analysis with an addition of fungal ITS analysis in Publication III. As using specific sequences means that we are not dealing with all the genes of all the available genomes from the sample sometimes it is argued that we should not use the term “metagenomics” but rather term like “metaprofiling” (Escobar-Zepeda et al. 2015).

A typical amplicon based study would include:

1) The selection of region of the genome/gene to amplify and suitable primers for such gene fragment. The selection of region depends on many factors. As an example 16S rDNA based analysis is considered. The choice of which 16S rDNA variable region to amplify depends on a variety of factors specific to the sample and experiment, including the particular bacteria present, whether it is most important to get resolution at a species, genus, or higher taxonomic level, and the gene fragment length that is afforded by the sequencer. Since all variable regions of the 16S rRNA gene have strengths and weaknesses, some researchers may opt to sequence more than one variable region to get a clearer view of the composition of the microbiome (Di Bella et al. 2013).
2) Gene fragment amplification and sequencing – (platform chosen by the considerations of budget vs. desired sequencing depth, the platform selection would also influence decisions for point 1 and vice-versa)
3) Quality-control and other processing of sequence data – the amplicon based methods are particularly sensitive to sequencing errors (Kunin et al. 2010), NGS platforms have different biases to produce different errors.

4) OTU clustering – needed to convert sequence information to "species/taxon" presence/abundance information. Matters considering OTU clustering will be covered extensively in the proceeding parts of this thesis.
5) OTU presence/abundance data can then be used as basis for different exploratory and statistical analyses.

The reliability of abundance data which is central in numerical ecology analyses is affected by multitude of aspects in all previous steps beginning with the amplicon selection (copy numbers differences/evolutionary conservation etc.) and culminating with OTU detection. Metaprofiling has been widely used due to its convenience to perform taxonomic and phylogenetic classification in large and complex samples within organisms from different life domains (Escobar-Zepeda et al. 2015). That enhances the suitability of the method for soil samples as soil microbial communities are one of the most complex. Soil-borne microbial communities are thought to be Earths greatest source of biodiversity, with estimates ranging from thousands to tens of thousands of species per gram of soil (Kowalchuk et al. 2007).

The use of specific amplicons as compared to shotgun metagenomics helps the economy of analysis both in getting better coverage for given sequencing depth and by reduced computational complexity. The choice of selecting amplicon length allows to consider different sequencing platforms to your experiment. For many possible amplicon regions there are databases assisting taxonomic and phylogenic classifications of sequences and OTUs produced. Databases governing 16S rDNA sequences being one of the most common. The examples of such databases are GreenGenes (DeSantis et al. 2006), Silva (Pruesse et al. 2007), and RDP (Maidak et al. 2001).

The popular 16S rDNA region amplicons have set of problems as highlighted by Escobar-Zepeda et al. (2015):
1) low resolution at the species level (Petrosino et al. 2009; Nalbantoglu et al. 2014)
2) a variable range in 16S rRNA gene copy number in many species (Acinas et al. 2004)
3) horizontal transfer of 16S rRNA genes (Schouls et al. 2003; Bodilis et al. 2012)
4) the fact that <0.1% of the total genome are ribosomal genes, hindering the amplification of this marker from very low abundant genomes in a sample.

Amplicon based approaches in themselves do not provide direct insight into the general functional state of microbial community but assembled phylogenetic and taxonomic information can to some extent cross-referenced with appropriate databases (De Filippo et al. 2012). All and all the popularity of amplicon based sequencing as compared to shotgun-metagenomics is easily understandable by the economical and ease of use aspects. Even if leaving these aspects out either of the methods can be more suitable for certain research questions.

3.2.3 Shotgun metagenomics

By comparison (to amplicon based strategies) in shotgun metagenomic analyses, all DNA in a sample is sequenced and analysed, which while being a more comprehensive approach means that the same sequencing depth yields far lower coverage in such studies. Acquiring the needed coverage puts pressure in the selection of sequencing platform. Computational requirements for such analysis will also be greatly increased. The assembly of longer sequences from a multitude of shotgun fragments is a daunting task especially when we have to account for the fact that different microbial taxa and genes may be present in the sample at wildly different abundances. All around the correct assembly of shotgun metagenomics assembly is still largely unresolved task although the number and quality of methods and tools is growing (Di Bella et al. 2013).

Some of the common steps in shotgun metagenomic analyses will be highlighted in following sections.

Binning refers to assigning reads to discrete "bins" on the account of common characteristics (sequence composition characteristics like codon usage, frequency of repeating elements etc.) on the assumption that such similar sequences might more likely be from same or similar organisms.

For gene finding and annotation it is more useful to have longer sequences – so often an assembly of overlapping sequences into longer contigs is in order. Such assembly might be done before and/or after the binning as both procedures can benefit from each other.

Assembly can be a purely de-novo process using only acquired sequences themselves and trying to find overlapping regions. Understandably process like this quickly demands ever larger resources with increasing number of sequences. Another approach is to map the sequences to reference database. While this reduces the computational needs it will also make the analysis depend on the reference database. There might be problems if the database contains errors or just has a very different composition of genes compared to sample, also it is not possible to map genes in the sample that are not represented in the database.

Whether the reads are assembled to longer contigs or not the next step would be **gene finding**. One possibility is to predict genes de-novo by the properties of the sequence. Examples of such tools are MetaGeneAnnotator (Noguchi et al. 2008), FragGeneScan (Rho et al. 2010), Prodigal (Hyatt et al. 2010), Orphelia (Hoff et al. 2009) and FragGeneScan + (Kim et al. 2015).

Afterwards the function of predicted genes could be determined by various tools and methods such as trying to map the gene to an appropriate database or using function prediction tools. A variety of databases and systems exist to aid in functional annotation of genes and gene segments, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 1999), the Clusters of Orthologous Groups (COG) system (Tatusov et al. 2003), Pfam (Bateman et al. 2004), the Conserved Domains Database (CDD) (Marchler-Bauer et al. 2005), SEED (Overbeek et al. 2005), TIGRFAM (Selengut et al. 2007) and eggNOG (Muller et al. 2009). These databases can find domains and

classify proteins by function, allowing determination of which functions and pathways are present in the metagenome or metatranscriptome, and in what abundances they are found.

Another approach is to immediately match read sequences against gene/protein databases, in this case the same database might already contain the necessary functional annotation therefore enabling the researcher to dispense with the function determining part. Signing oneself to only this approach leaves out possibility of finding novel genes and predict their function though.

To compare samples by found genes and their abundances a multitude of tools is also available. Jonsson and co-workers (2016) compared 14 different tools for identification of differentially abundant genes between metagenomes and found edgeR (Matsen and Evans 2013) and DESeq2 (Love et al. 2014) to have overall best performance.

In conclusion: shotgun metagenomics allows many additional options compared to amplicon-based strategies, but at the cost of increased complexities – some of which might turn out prohibitive for given research goal.

3.3. Data analysis approaches in amplicon based soil metagenomics studies

3.3.1. General overview of clustering methods used in 16S rDNA analyses

Although there are alternatives, especially when the main goal of our numerical ecology analysis is centered on sample comparison, one of the main methods to transform metagenomic sequence data to data more ecologically meaningful is through the process called clustering. Similar sequences are clustered together into OTU-s (Operational Taxonomic Units) that alongside their abundance information can be used as a stand-in for species (or other taxonomic level grouping) in already existing ecological methodologies. From OTU data it is possible to calculate the richness and diversity indices, compare samples using multivariate statistical analysis, correlations of OTU-s with parameters measured etc. OTU clustering methods could be classified in three general categories – hierarchical clustering, heuristic clustering and model-based clustering methods (Chen et al. 2013).

In the hierarchical clustering category, a distance matrix measuring the difference between each pair of sequences is calculated first, and standard hierarchical clustering is then used to define OTUs at a specific level of sequence dissimilarity. Most of these methods have an $O(N^2)$ computational complexity, where N is the number of sequences, posing a significant computational bottleneck for processing large-scale sequencing datasets (Chen et al. 2013). Different heuristic algorithms have been developed mainly to avoid the computational bottlenecks that might be prohibitive when analysing large scale datasets. A usual way to reduce the quadratic complexity is to ditch the computation of full

pairwise distance matrix and use algorithms that would process input sequences and distance calculations sequentially. Examples of such algorithms are CD-Hit (Li and Godzik 2006) and UCLUST (Edgar 2010).

Both these methods use pairwise sequence alignment and process input sequences sequentially. Given a predefined threshold, an input sequence is either assigned to an existing cluster if the distance between the sequence and a seed is smaller than the threshold, or becomes a seed otherwise. The computational complexity of greedy heuristic clustering is on the order of O(NM), where M is the number of seeds and usually M << N (Cai and Sun 2011).

Greedy heuristic clustering (e.g. CD-HIT and UCLUST) processes input sequences one at a time, avoiding the expensive step of comparing all pairs of sequences. Given a predefined threshold, an input sequence is either assigned to an existing cluster if the distance between the sequence and a seed (the sequence representing that cluster) is smaller than the threshold, or becomes a new seed for a new cluster otherwise. Consequently, the computational complexity of greedy heuristic clustering is O(MN), where M is the number of seeds. Usually M<<N, and hence greedy heuristic clustering is computationally much more efficient than HC (Sun et al. 2012).

These greedy clustering methods suffer from two fundamental problems. Mahé et al. (2014) summarizes them as following.

First, they use an arbitrary fixed global clustering threshold. As lineages evolve at variable rates, no single cut-off value can accommodate the entire tree of life. A single global clustering threshold will inevitably be too relaxed for slow-evolving lineages and too stringent for rapidly evolving ones (Stackebrandt and Goebel 1994; Sogin et al. 2006; Nebel et al. 2011; Koeppel and Wu 2013). Secondly, the input order of amplicons strongly influences the clustering results. Previous centroid selections are not re-evaluated as clustering progresses, which can generate inaccurately formed OTUs, where closely related amplicons can be separated and unrelated amplicons can be grouped (Koeppel and Wu 2013).

These two problems are brought out as a rationale for the development of Swarm algorithm (Mahé et al. 2014) which can be according to Chen et. al. (2013) classification still be described as hierarchical algorithm though it has some specific properties not common with most hierarchical methods as described by Chen et. al., (2013) and which will be covered more thoroughly in separate chapter dedicated to Swarm algorithm/tool. The possibility to avoid these kind of arbitrary cut-offs is also one rationale to development of CROP (Hao et al. 2011) algorithm which is an unsupervised Bayesian clustering algorithm using Gaussian mixture models. It will also be discussed in separate chapter and leads us to the third class of OTU clustering methods.

A third class of OTU clustering methods as classified by Chen et al. are model based OTU clustering methods. Probability models have been proposed for quite some time as a basis for cluster analysis. In this approach, the data are viewed as coming from a mixture of probability distributions, each representing a different cluster (Fraley and Raftery 1998).

Clustering algorithms based on probability models offer a principle alternative to heuristic algorithms. In particular, model-based clustering assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. The issues of selecting a 'good' clustering method and determining the 'correct' number of clusters are reduced to model selection problems in the probability framework. Gaussian mixture models have been shown to be a powerful tool for clustering in many applications (Yeung et al. 2001).

As for the meaning of "quite some time" – in the article (Fraley and Raftery 2002) declare that mixture based models have often been proposed and studied in the context of clustering – citing sources that go back to 1960's. Therefore it comes as no surprise that such mature methods, which have shown their uses in many different clustering applications, have now found their applications in OTU clustering tools like CROP or BEBaC (Cheng et al. 2012).

3.3.2. Mothur software package

DOTUR (Schloss and Handelsman 2005) was introduced in 2005 as a software tool that both did OTU assignment of sequences by the means of hierarchical clustering based on PHYLIP (Felsenstein 1989) generated sequence distance matrices (using nearest, furthest or average neighbour joining algorithm as chosen by user). Mothur, version 1.0.0 released in February 2009, is a software platform that incorporates algorithms from many previous tools such as DOTUR, SONS (Schloss and Handelsman 2006a), TreeClimber (Schloss and Handelsman 2006b), LIBSHUFF (Singleton et al. 2001), ∫-LIBSHUFF (Schloss et al. 2004), and UniFrac (Lozupone and Knight 2005). The platform has been in continual development with several new version releases per year, it has continued to grow in capabilities and different algorithms implemented from many other bioinformatics tools.

The capabilities of first version already included over 25 calculators for quantifying key ecological parameters for measuring α- and β-diversity; visualization tools including Venn diagrams, heat maps, and dendrograms, functions for screening sequence collections based on quality; a NAST-based sequence aligner; a pairwise sequence distance calculator; and the ability to either call individual commands from within Mothur, using files with lists of commands (i.e. batch files), or directly from the command line which provide for greater flexibility in setting up analysis pipelines (Schloss et al. 2009).

The concise way to describe Mothur cluster generation used in this work and in Publication I would be to say that we used average neighbour algorithm for clustering sparse distance matrix (distances larger than 0.1 discarded) generated from multiple sequence alignment by simple pairwise comparison (gaps penalised as one difference). All the steps (alignment, distance calculation, clustering) were done within Mothur package. More details are given in the Materials and methods section.

Preem et al. (2012) used Mothur v.1.13.0 for sequence data pre-processing, distance matrix generation, clustering, ecological index calculation and more.

In current work all other data sets were analysed using Mothur package v.1.36.0. to some extent – pre- and post-clustering. Clustering itself is done by different tools for such analyses. The data pre-processing is covered in Materials and methods section.

3.3.3. UCLUST and USEARCH

UCLUST is an example of heuristic clustering algorithm. As described earlier in general discussion it achieves linear complexity by the means of processing input sequences one-by-one by assigning them to (randomly seeded) OTU-s by pre-defined threshold. It uses USEARCH algorithm for the sequence assignment. A brief description of its rationale is given in the following paragraph.

High-throughput is achieved by using a fast heuristic designed to enable rapid identification of one or a few good hits rather than all homologous sequences. For a given query, database sequences are sorted in order of decreasing number of words in common to exploit the fact that similar sequences tend to have short words in common (see e.g. (Edgar 2004)). When examined in this order (i) if a hit exists in the database, it is likely to be found among the first few candidates, and (ii) the probability that a hit exists falls rapidly as the number of failed attempts increases. A search can therefore often be terminated after examining a small number of candidates without a large cost in sensitivity (Edgar 2010).

The UCLUST algorithm employs USEARCH to seek a matching cluster for a given sequence. An initially empty database is created, which is extended as input sequences are processed. The database contains exactly one representative sequence for each cluster, known as its seed. Each input sequence (Q) is compared to the current database using USEARCH. If a matching seed is found, Q is assigned to the corresponding cluster and the database remains unchanged; otherwise, Q is added to the database, becoming the seed of a new cluster. A match is defined as a global alignment of the query to a seed with an identity exceeding a pre-set threshold. Using a single representative sequence per cluster minimizes the database size and total number of sequence comparisons and hence the time and memory required, but may not be biologically optimal as there is no enforced lower bound on pair wise similarity when neither sequence is a seed (from supplemental materials of (Edgar 2010), online version is available at http://bioinformatics.oxfordjournals.org/content/suppl/2010/08/11/btq461.DC1/supp_mat_rev2.pdf)

3.3.4. CROP – an unsupervised Bayesian clustering method

CROP is unsupervised Bayesian clustering algorithm that uses Gaussian mixture model to describe 16S rDNA sequence data as such it falls in the model-based camp of OTU clustering methods as classified by Chen et al. (2013) . The main rationale for the CROP clustering method as described in the Hao et al. (2011) by authors themselves is the sensitivity to dissimilarity threshold and the difficulty of choosing an optimal threshold especially in the light of sequencing errors that crop up as problems in traditional hierarchical clustering methods.

Article by Hao and co-workers (2011) brings out many references to studies that show the overestimation of OTU-s by hierarchical clustering methods (Quince et al. 2009; Marco 2010; Huse et al. 2010). Also they highlight that distance calculated from multiple alignment was, in general, larger than those calculated from pairwise alignments and thus might also result in overestimation of the number of OTUs (Sun et al. 2009).

A concise summary of CROP can be found in publication by Soueidan & Nikolski (2015). CROP (Hao et al. 2011) (Hao et al. 2011) (Hao et al. 2011) (Hao et al. 2011) uses an unsupervised Bayesian clustering approach. The model relies on the notion of probability that a given sequence s belongs to a cluster. This probability is defined as function of the distance between the sequences and the sequence that is in the center of the cluster. Moreover, CROP applies the divide-and-conquer principle by dividing the dataset into small subsets and performing Bayesian clustering on the subsets. Thus generated clusters are replaced by their consensus sequences on which a final step of Bayesian clustering is performed in order to obtain the OTUs.

For a more elaborate summary of CROP principles the following paragraphs as written by Hao et al. (2011) authors themselves should be suitable- “The key concept of our method replaces the mean value of a Gaussian distribution and instead uses a ‘center’ sequence to characterize a specific cluster. Thus, if we consider the sequences as data points in a high-dimensional space and we calculate the pairwise distances as the distance between two data points, then the probability that a sequence belongs to a cluster becomes a function of the distance between the sequence and the center. The nature of Gaussian distributions can handle sequencing errors as well as sequence variations. However, by restricting the parameter space of the standard deviations of the Gaussian distributions, we could limit our probabilistic search to the parameter subspace in which the clustering results reflect the desired partitions of the datasets and, hence, the accurate number of underlying OTUs.

Based on this model, we can define the likelihood of the data and use a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior distribution of the parameters to obtain the optimal clustering. The optimal result, which maximizes the posterior probability, will give all the quantities of interest, including the number of clusters, their relative abundance levels and the sequences in each cluster. Richardson and Green (1997) and Stephens (2000) have proposed MCMC methods to study the mixture model with an

unknown number of components. In this application, we used a Markov birth–death process to build the Markov Chain with appropriate stationary distribution, as proposed by Stephens (2000). That is, in each step, a new cluster would be created, or an existing one would be deleted, according to which operation is more likely to increase the posterior probability. To enhance computational efficiency, we further introduced a hierarchical approach by splitting the data into small blocks, running Bayesian Clustering on each block independently, and then later merging these clustering results. We also introduced several criteria to reduce the burden of calculating Gaussian density functions, thereby accelerating the MCMC process".

3.3.5. Swarm

Swarm could be described as agglomerative hierarchical clustering algorithm. It's rationale as described by authors is to create an exact clustering algorithm that could compete with greedy heuristic based algorithms in terms of computational complexity and speed.

Current chapter describes Swarm on the basis of Mahé et al. 2014. The main assumption of Swarm is that if amplicons are viewed as discrete points in abstract amplicon-space then the clusters forming from directly neighbouring amplicons (one nucleotide difference) are separated by each other by empty regions, meaning that amplicons do not form a vast uninterrupted continuum. If that assumption holds true OTU-s can be allowed to grow iteratively until they reach their natural limits – empty space around amplicon clusters.

Swarm explores the amplicon-space as follows: Swarm processes the input file and creates a pool of amplicons. An empty OTU is created, and the first available amplicon in the pool is withdrawn from the pool to become the OTU seed. The seed is then compared to all amplicons remaining in the pool, and the measured number of differences is stored in the memory. The number of differences is calculated as the number of nucleotide mismatches (substitution, insertion, or deletion) between two amplicons once the optimal pairwise global alignment has been found. Amplicons for which the number of differences is equal to or less than d, the user-chosen local clustering threshold, are removed from the pool and added to the OTU where they become subseeds. Each subseed is then compared to the amplicons remaining in the pool, but only to those that have at most 2d differences with the seed. Indeed, amplicons with more than 2d differences with the seed cannot have d or less differences with one of its subseeds.

This iterative growth process is repeated for each generation of subseeds as long as new amplicons are captured. The OTU is then closed. The first available amplicon is removed from the pool, becomes the seed of a new OTU, and the process is repeated until no more amplicons remain in the pool. This clustering process generates stable OTUs, regardless of the first seed choice. Thus, an OTU organically grows to its natural limits where it cannot recruit any more

amplicons with d or fewer differences. Operating in this way, Swarm removes the two main sources of variability inherent in greedy de novo clustering methods: the need to designate an OTU centre (centroid selection), and the need for an arbitrary global clustering threshold (maximum radius). Swarm outlines OTUs without imposing one particular shape or size, and produces the same OTUs regardless of the initially selected amplicon.

Under certain conditions (like using short and/or slowly evolving markers), the assumption that amplicons do not form a vast continuum can be violated. Huse et al. 2010 have shown that single-linkage clustering is known to produce chains of amplicons that can potentially link closely related OTUs and decrease clustering resolution. To solve this issue, Swarm implements a breaking phase that uses the structure of the cluster and amplicon abundance values to eliminate weak contiguity regions, and to delineate higher-resolution OTUs.

Swarm internally produces a graph representation of the OTU, in the form of a star-shaped minimum spanning tree. OTUs present an internal structure where the most abundant amplicon usually occupies a central position and is surrounded by less abundant amplicons. To identify and break the chains, our algorithm finds paths linking abundant amplicons (peaks) and monitors the abundance variations along these paths. If abundances decrease, go through a minimum, and then go up again (valley shape), this indicates a possible amplicon chain. Depending on the depth of the valley (i.e., the ratio between the minimum and maximum observed abundances), the algorithm will decide whether or not to break the graph into independent OTUs.

Concerned with possible under-grouping in the original Swarm algorithm a new version was published which includes so called fastidious option for grafting singletons and doubletons to larger OTU-s (Mahé et al. 2015). In this work we use Swarm both with and without fastidious option.

4. MATERIALS AND METHODS

4.1. Description of study sites and sampling

For all data sets of OTU based analyses the same procedure as in the Publication I described applies. Five soil samples (50 cm^3 of each) were collected from a 0 to 10 cm layer of the soil A horizon. A 10-meter-wide square strategy was applied for this purpose on each study site. Four samples were collected from the corners and one from the middle point of the square. The roots were separated from the soil. 10 g of each soil sample was stored at −20°C for DNA extraction, and the rest was stored at 4°C for chemical analyses. Soil pH_{KCl}, total nitrogen, phosphorus (ammonium lactate extractable) and the amount of organic matter (LOI – loss on ignition) were determined for each sample. In addition, the elemental composition (concentrations of total Al, B, Ca, Cd, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb and Zn) of each soil was determined by inductively coupled Plasma Mass Spectrometry (ICP, wet digestion with HNO 3 –H 2 O 2) in the laboratory of the Finnish Forest Research Institute (Vantaa, Finland).

4.2. DNA extraction, PCR product preparation and sequencing

For all sets of OTU based analyses the same procedure as in the publication (Preem et al., 2012) described applies. Total DNA was extracted from soil using an UltraClean Soil DNA kit (Mo Bio Laboratories, Inc.) according to the manufacturer's protocol. PCR was carried out using Qiagen HotStarTaq mix (QiagenGmBh, Germany) and primers 8F (Edwards et al. 1989) and 357R (Muyzer et al. 1993) amplifying V2 and partly V3 hyper-variable region of the 16S rRNA gene. The forward 8F primers (5-GCCTCCCTCGCGCCATCAG (NNNNNN) AGAGTTTGATCCTGGCTCAG-3) used in the reaction mixture contained sequencing primer (underlined) followed by a 6-bp barcode (indicated in parentheses) unique to each sample (Parameswaran et al. 2007) and finally the primer sequence. The reverse 357R primer (5-GCCTTGCCAGCCCGCTCAG CTGCTG CCTCCCGTAGG-3) had a sequencing primer (underlined) before the primer sequence. The following PCR program was used: 15 min at 95°C followed by 3 cycles of 30 s at 95°C, 30 s at 50°C and 60 s at 72°C and 28 cycles of 30 s at 95°C, 30 s at 65°C and 60 s at 72°C and final extension for 10 min at 72°C. PCR products were purified from gel using the Qiagen QIAquick Gel Extraction Kit (Qiagen Gmbh,Germany), quantified using NanoDrop 1000 (Thermo Scientific) and finally mixed at equimolar concentrations prior to sequencing. DNA extraction and PCR products preparation for sequencing was performed by BiotaP LLC (Tallinn, Estonia). Sequencing was performed by GATC Biotech (Constanz, Germany) using a Genome Sequencer FLX System (Roche Applied Science).

4.3. Data analysis

4.3.1. Sequence data pre-processing

For all sets of OTU based analyses the same procedure for sequence data pre-processing as described in the Publication I, including homopolymer removal, removal of short sequences and chimeras, applies. The following data denoising procedure was performed: all sequences containing any ambiguous bases and sequences containing long stretches of homopolymers (more than 7 bp) were removed (Margulies et al. 2005), sequences less than 180 bp were removed and identical sequences were merged. To remove possible chimeric sequences we used UCHIME de novo approach (without reference database) as an incorporated command in Mothur (Edgar et al. 2011).

4.3.2. OTU clustering

For Mothur based analysis following procedure was used. The pre-processed dereplicated sequences were aligned and then pairwise distances were calculated using Mothur "dist.seqs" command (simple pairwise comparison of aligned sequences, consecutive gaps in sequence penalized as one gap, gaps at the end of sequence penalized, distances larger than 0.1 discarded from result). Thereafter OTU-s were created by Mothur command "cluster" using average neighbour algorithm. Version of Mothur used was v.1.34.4. The process is also described in Publication sub-section "2.3. Sequenced data processing and taxonomic assessment", in this case v.1.13.0. of Mothur was used.

For CROP based analysis the pre-processed redundant sequence set was used as input for CROP program to perform the clustering at species level (CROP parameter "-s").

For UCLUST based analysis part of QIIME (Caporaso et al. 2010) pipeline were used. The pre-processed redundant fasta sequences were supplied with QIIME compatible labels by add_qiime_labels.py QIIME script after which the QIIME script pick_otus.py was used which by default uses UCLUST algorithm for clustering.

For Swarm based analysis parameter value of **d=1** (maximum allowed differences between amplicons is 1 mismatch) and -f (fastidious) options were used. The input was dereplicated pre-processed sequence set with numbers of sequence occurrence added to the sequence identifiers by in house scripts.

4.3.3. OTU data post-processing and statistical analyses

As a first step before calculating any statistics from acquired OTU data – OTU-s containing only one single sequence were culled. The removal was performed for CROP, Swarm and UCLUST based analyses, and also for analysis that was otherwise identical to the Publication I which will be mentioned as Mothur based analysis from now onward.

Owing to the differences in OTU clustering and subsequent removal of singletons the sequence data for Mothur based analysis was normalized to 2437 sequences per sample and 18045 sequences per site. The sequencing data for CROP analysis was normalized to 2558 sequences per sample and 18714 sequences per site, for UCLUST analysis it was normalized to 2483 sequences per sample and 18252 sequences per site and for Swarm analysis 360 sequences per sample and 3320 sequences per site. Nonparametric Kruskal–Wallis one-way analysis of variance by ranks was applied to verify differences in diversity indices and the relative abundance of bacterial phylogenetic groups between the studied locations.

The number of OTU-s and diversity indices acquired by different methods was compared by the use repeated measures ANOVA. Post-hoc testing was done by multiple t-tests using Holm-Bonferroni correction.

For each set of analyses Bray-Curtis dissimilarity measure was applied to calculate the between samples distance matrices. The Congruence Among Distance Matrices (CADM) analysis (Legendre et al. 2004) was performed to find congruence among these distance matrices. The distance matrices were also compared by means of Mantel tests. Principal coordinate analysis (PCoA) was used to explore and visualize similarities in a low-dimensional space between soil samples based on the obtained distance matrices. Based on the same distance matrices also one-way permutational multivariate analysis (PERMANOVA) (Anderson 2001) was performed to test for differences in microbial community composition between the studied locations. Before conducting PERMANOVA, the distance-based test for the homogeneity of multivariate dispersions was performed. Also distance-based regression analysis was applied to identify variables that explained significant amounts of variation in bacterial community structure. The analysis was carried out using the DISTLM (McArdle and Anderson 2001) program with forward selection procedure and 10000 permutations.

Molecular Ecological Network Analyses Pipeline (MENAP) was used to create molecular ecological networks from the obtained OTUs (Deng et al. 2012). Input for the MENAP pipeline were post-processed OTU abundance tables gathered from every analysis. Only OTU-s that were present at least in 5 samples were included in the network analysis.

4.4. Mock community generation and analysis

To evaluate different OTU clustering methods against a dataset with sequences of precisely known origin a low-complexity mock dataset was constructed. For this sequences which had perfect matches for primers used in Publication I (8F, 357R) were acquired from Greengenes (DeSantis et al. 2006) database (2013-August version). Further on, a hundred random sequences that had Greengenes classifications to species level were chosen such that each one represented a different species. These 100 unique sequences were used to construct a dataset

consisting of 114 744 sequences by using a simulation tool called Grinder (Angly et al. 2012). Abundance model by which the unique sequences were multiplied was following – a random Gaussian distribution was generated with a median of 1000 and standard deviation of 200, acquired numbers were rounded to integers and used as count numbers for OTUs. The error model used for this mock dataset generation was 454 pyrosequencing error model as described in (Balzer et al. 2011).

The resulting dataset was then submitted to the same set of analyses that were performed previously in this work on the Publication I dataset. Including similar sequence data pre- and post-processing and OTU clustering by four different OTU clustering methods. Following OTU clustering, to assess the quality of generated OTU data, the acquired OTU abundance profiles were compared to abundance profile used when generating the mock dataset – by means of bootstrapped Kolmogorov-Smirnov tests. Also for that aim statistics such as number of OTU-s observed, ecological indices and the representative sequence for 10 most abundant OTUs were acquired.

4.5. Clustering independent analysis

In order to provide a clustering independent comparison analysis we took following steps. As described in (Steven N. Evans, Matsen et al. 2012) the widely used weighted Unifrac distance is just Kantorovich-Rubenstein distance if we equate a metagenomic sample with its empirical distribution on a reference phylogenetic tree. We took the samples from the original Publications I-III and used the SEPP tool (Mirarab et al.) to produce such placements onto the Greengenes reference tree. Kantorovich-Rubenstein distances between samples were calculated from phylogenetic placements using guppy tool from pplacer suite (Matsen et al. 2010). Several PCoA-s based ordination plots on such pairwise distances were produced containing different combinations of datasets. The PCoA-s (especially those containing Publication I dataset) provide an OTU clustering independent analysis that can be used for the background.

5. RESULTS AND DISCUSSION

5.1. Differences in the quality of estimating the mock community structure

The mock community was generated form 100 unique sequences which were multiplied while allocating a certain amount of sequence differences (amounting to 454 pyrosequencing error probabilities). The assumption being that each unique sequence should in this way represent an OTU center – ideally with its "progeny" sequences (differentiating from parent by errors introduced in multiplication process) clustering in the same OTU. **Table 1** shows the number of OTU-s observed and calculated inverse Simpson indices for mock dataset and for analyses after clustering the dataset with four different clustering methods. It can be seen that if we take the original amount of unique sequences (100) as the actual amount of OTU-s in dataset then most of the clustering methods result in overestimation of OTU numbers compared to mock data.

Table 1. OTU-s observed and inverted Simpson's indices for mock OTU data and OTU- data created by different clustering methods.

	Mock	Mothur	CROP	UCLUST	Swarm	Swarm d=4
Number of OTU-s	100	278	93	204	493	96
Inverted Simpson's index	96,8	91,0	76,5	82,9	123,1	76.3

Particularly large overestimation was for the case of Swarm analysis using the default parameter for maximum differences for amplicons to be grouped together **d=1** and the fastidious option turned on. As the generation of OTU-s in Swarm algorithm goes agglomeratively by adding amplicons that differ by d differences to each other one-by-one until no such amplicon can be found. The most organic **d** especially if aiming for a high taxonomical resolution seems **d=1.** Fastidious option merges OTU-s by postulating "virtual amplicons" which if existing would allow merger of small OTU-s (singletons and doubletons by default) to be merged with some other and then doing the merges accordingly. Such process makes a good deal of biological sense as no matter how different a set of amplicons is if it is possible to link them a one difference step by one difference step at a time representing phylogenetically linked amplicons where each is separated by other by a mutational event.

Unfortunately, for Swarm algorithm our mock dataset is constructed by adding 454 pyrosequencing errors to a seed amplicon – so the "progeny" might not be removed from seed and each other in such step-by-step evolution. So it

seems understandable that such small difference as **d=1** will create much more OTU-s as the "progeny" sequences can often both be removed from seed sequence by a larger amount of differences and there is no step-by-step sequence of mutation events that would lead from seed through one progeny sequence to another. To alleviate such problems we did some tests for clustering with Swarm using larger d values.

At **d** value 4, Swarm clustered 98 OTU-s (result of 96 OTUs observed for our analysis where we removed 2 singletons) which was closest to original set of 100 seed sequences. The Mothur and UCLUST analyses ended up overestimating the number of OTU-s by a smaller amount and CROP analysis yielded 93 OTU-s – close to the original number of seed sequences.

The diversity as expressed by inverted Simpson index shows that although Swarm **d=4** and CROP analysis result in a number of observed OTU-s pretty close to the original seed sequences the abundance profiles are still not similar to original mock dataset being on the whole less diverse – even Mothur analysis that has 204 observed OTU-s as a result still has an inverse Simpson index that shows a lower diversity than in the original dataset.

OTU abundances for each set of analysis were compared to the original set by means of bootstrapped (10000 repeats) Kolmogorov-Smirnov tests. At p-values much lower than 0.001 none of these could be considered as coming from the same distribution as original.

In **Table 2** the abundances and representative sequences of 10 most abundant OTU-s are shown. It is interesting to see how the most abundant OTU for Mothur, CROP and Swarm **d=1** analysis all have the same representative sequence – sequence 131 which is based from an original sequence (from the selection of 100 unique seed sequences) 470487 – *Helicobacter canadensis*. Suggesting similarities in the OTU generating process. The Swarm **d=4** has the most abundant OTU representative sequence as 14312 based on original seed sequence of 274316 *Helicobacter hepaticus*. In original mock set the seed sequence that was amplified the most was the sequence 11111417 – *Pelomonas puraquae*.

Table 2. Classification and abundance of 10 most abundant OTU-s by OTU clustering method. Shown are the classifications of the OTU cluster representative/center/seed sequence and amount of times a sequence belonging to certain OTU has been observed in samples. Also the ID of the cluster center sequence is shown for comparison purposes.

Mock sequences					
counts	**seqid**	**classification**	**counts**	**seqid**	**classification**
1595	1111417	*Pelomonas puraquae*	1498	1015954	*Cylindrospermopsis raciborskii*
1550	1109455	*Octadecabacter antarcticus*	1495	781232	*Thalassobacter stenotrophicus*
1543	1107772	*Chromatium okenii*	1487	1056070	*Hyphomicrobium zavarzinii*
1509	808546	*Rhodopila globiformis*	1474	825031	*Methylobacterium adhaesivum*
1506	801294	*Brevundimonas diminuta*	1443	638485	*Fibrobacter succinogenes*
Mothur			**CROP**		
counts	**seqid**	**classification**	**counts**	**seqid**	**classification**
2272	131 ref.470487	*Helicobacter canadensis*	3104	131 ref.470487	*Helicobacter canadensis*
2187	125 ref.801294	*Brevundimonas diminuta*	2208	90 ref.576765	*Brevundimonas vesicularis*
1228	646 ref.224417	*Rhizobium leguminosarum*	1874	749 ref.781232	*Thalassobacter stenotrophicus*
1221	49 ref.163757	*Roseobacter denitrificans*	1855	908 ref.240979	*Pseudoruegeria indica*
1174	477 ref.251262	*Methylobacterium organophilum*	1726	9130 ref.189350	*Bartonella bacilliformis*
1149	18 ref.169604	*Sphingobium estrogenivorans*	1715	54 ref.155996	*Bradyrhizobium elkanii*
1104	409 ref.251377	*Anabaena cylindrica*	1384	272 ref.565851	*Thiodictyon syntrophicum*
1100	1277 ref.1063320	*Anaerospora hongkongensis*	1251	646 ref.224417	*Rhizobium leguminosarum*
1065	173 ref.346882	*Paracoccus aminovorans*	1221	49 ref.163757	*Roseobacter denitrificans*
1060	13 ref.73433	*Nannocystis exedens*	1207	477 ref.251262	*Methylobacterium organophilum*

UCLUST			Swarm		
counts	**seqid**	**classification**	**counts**	**seqid**	**classification**
2800	131 ref.470487	*Helicobacter canadensis*	124	131 ref.470487	*Helicobacter canadensis*
2149	125 ref.801294	*Brevundimonas diminuta*	118	323 ref.140774	*Pseudoalteromonas luteoviolacea*
1660	908 ref.240979	*Pseudoruegeria indica*	112	560 ref.129026	*Ruegeria pomeroyi*
1608	13114 ref.347498	*Clostridium hiranonis*	110	451 ref.1107772	*Chromatium okenii*
1515	388 ref.111946	*Azorhizobium doebereinerae*	109	388 ref.111946	*Azorhizobium doebereinerae*
1251	93039 ref.342971	*Gluconacetobacter diazotrophicus*	109	56 ref.532163	*Rhodobacter sphaeroides*
1219	29915 ref.136508	*Candidatus Liberibacter americanus*	104	646 ref.224417	*Rhizobium leguminosarum*
1198	109652 ref.247213	*Methylosinus sporium*	103	464 ref.136508	*Candidatus Liberibacter americanus*
1169	18 ref.169604	*Sphingobium estrogenivorans*	101	17 ref.327793	*Loktanella vestfoldensis*
1108	409 ref.251377	*Anabaena cylindrica*	99	1299 ref.549334	*Paracoccus zeaxanthinifaciens*

5.2. Differences in evaluated bacterial community structure

The number of OTU-s obtained did not differ between sites for analyses using any clustering methods (Kruskal-Wallis test, $p>0.05$). The calculated Inverted Simpson's diversity indices showed no difference between the studied sites (Kruskal–Wallis test, $p > 0.05$) in any sets of analyses. To visualize the differences between the studied sites according to the numbers of obtained OTUs, four-way Venn diagrams were created for each sets of analyses. These Venn diagrams can be seen in Figure 1.

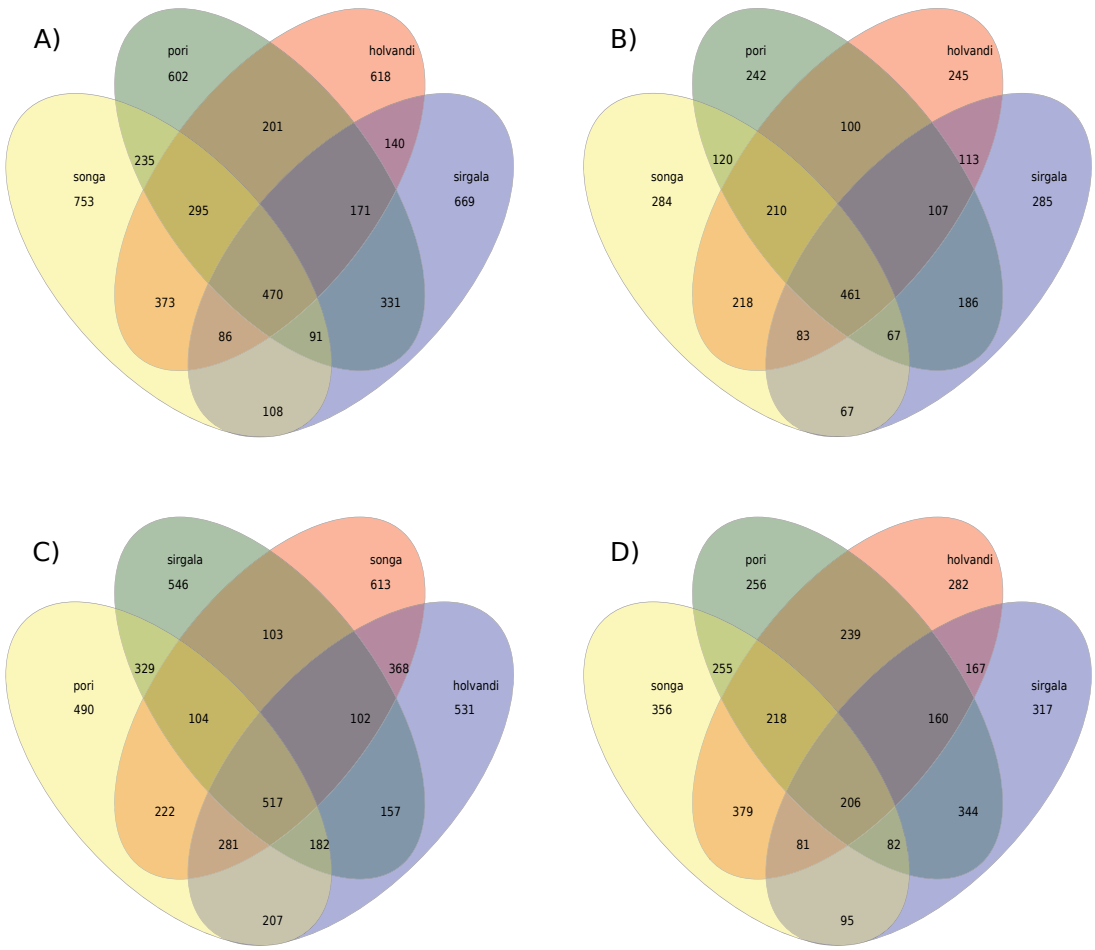


Figure 1. Unique and shared OTU-s by the sampling sites according to the results of different OTU clustering methods. Venn diagrams illustrate the OTU-s found in four different alder stands (Holvandi, Porijõgi, Sirgala, Songa). The diagrams are labeled such that A, B, C, D correspond to Mothur, CROP, UCLUST and Swarm clustering methods, respectively.

The total number of observed OTU-s was highest in Mothur based set with 5143 OTU-s observed. The number of OTU-s observed was 4752 for Swarm based set, 3437 for UCLUST based set, and 2788 for CROP based set. The percentage of OTU-s unique to stands was highest in UCLUST based set – 63.4 % of sequences were not shared between the stands. This percentage was 51.4 in Mothur based set, 37.9 in CROP based set and 25.5 in Swarm based set. The percentage of OTU-s shared by all stands was highest in CROP based set – 16.5% of all OTU-s were shared between all stands. This percentage was 15.0 for UCLUST based set, 9.1 for Mothur based set and 4.3 for Swarm based set.
The proportions of OTU-s shared by all stands were statistically dependent on clustering methods used (Fisher's exact test – $p<0.001$) the same applies to the proportions of unique OTU-s (Fischer's exact test – $p<0.001$). The removal of singletons seems to mostly have affected the OTU-s that were unique – in comparison compare in 2012 Mothur based analysis over 71% of sequences were unique between the stands.

The overall OTU numbers and IS indices were found to be significantly different between different clustering approaches by the means of repeated measures ANOVA ($p<0.001$ in both cases). The results from post-hoc tests are depicted in Table 3**.**

While the PERMANOVA analysis found that the communities of the stands are statistically different ($p<0.01$) for the Mothur based analysis in Publication I, for the current tests there was no statistically significant difference between the stands in any of the four sets of analyses. Similar to the Publication I there were no significant differences in multivariate dispersions between the four studied stands ($p > 0.05$). The loss of statistically significant differences between the stands for Mothur based set should be accounted for the removal of singleton OTU-s.

CADM analysis showed Bray-Curtis distance matrices acquired for different sets of analyses were highly congruent (Kendalls coefficient of concordance $W =0.94$, permutational probability calculated by 10000 permutations $p<0.001$). Mantel tests were also conducted between the Bray-Curtis distance matrices from different sets. According to the Mantel tests the Swarm set of analyses produced a Bray-Curtis dissimilarity matrix that compared to any other matrices produced less similarities than any other two-way comparison (**Table 4).**

Table 3. Results from paired t-tests on OTU numbers (S_{obs}) and IS indices acquired by analysing OTU data produced by different OTU clustering methods. Shown are the (Holm-Bonferroni corrected) p-values for each possible pairwise combination.

S_{obs}	Mothur	CROP	UCLUST
Mothur			
CROP	0.001		
UCLUST	0.093	0.001	
Swarm	0.001	0.037	0.0002
IS indices			
Mothur			
CROP	0.001		
UCLUST	0.093	0.001	
Swarm	0.001	0.038	0.002

Table 4. Mantel correlations, computed on rank-transformed distances. Distances were acquired from Bray-Curtis distance matrices based on OTU data gathered using different OTU clustering method on same samples. Correlations shown are Pearson correlations on rank transformed variables A.K.A Spearman correlations. Also Kendall's coefficient of concordance, W = 0.94 – was calculated for the set. Holm-Bonferroni corrected permutational p-values for all coefficients were $p<0.001$. Tests were carried out using 10,000 permutations.

Spearman correlations	Mothur	CROP	UCLUST	Swarm
Mothur	1	0.98	0.99	0.85
CROP	0.98	1	0.98	0.83
UCLUST	0.99	0.98	1	0.84
Swarm	0.85	0.83	0.84	1

The same Bray-Curtis distance matrices were also used in principal coordinate analyses – the results of can be seen in **Figure 2.** Obtained PCoA plots indicate that the general sample placement remains quite similar no matter the clustering method. The biggest visual difference being again the Swarm method result.

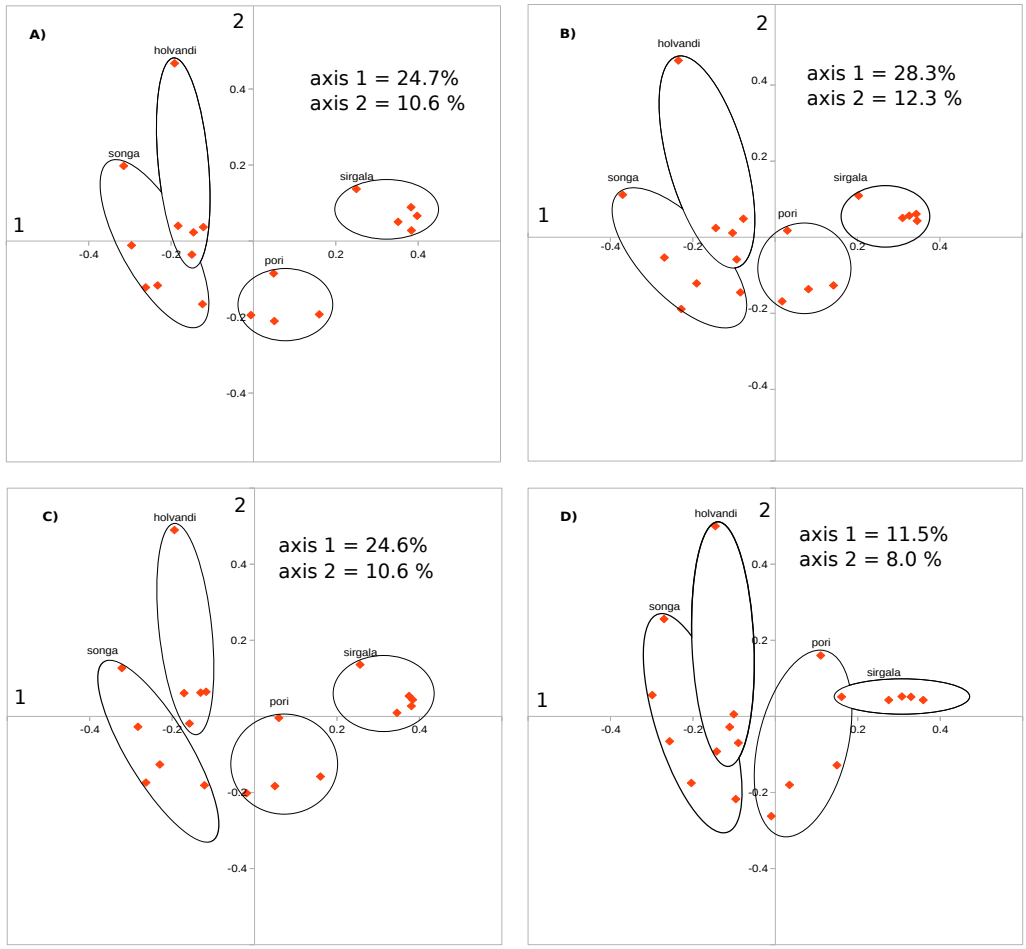
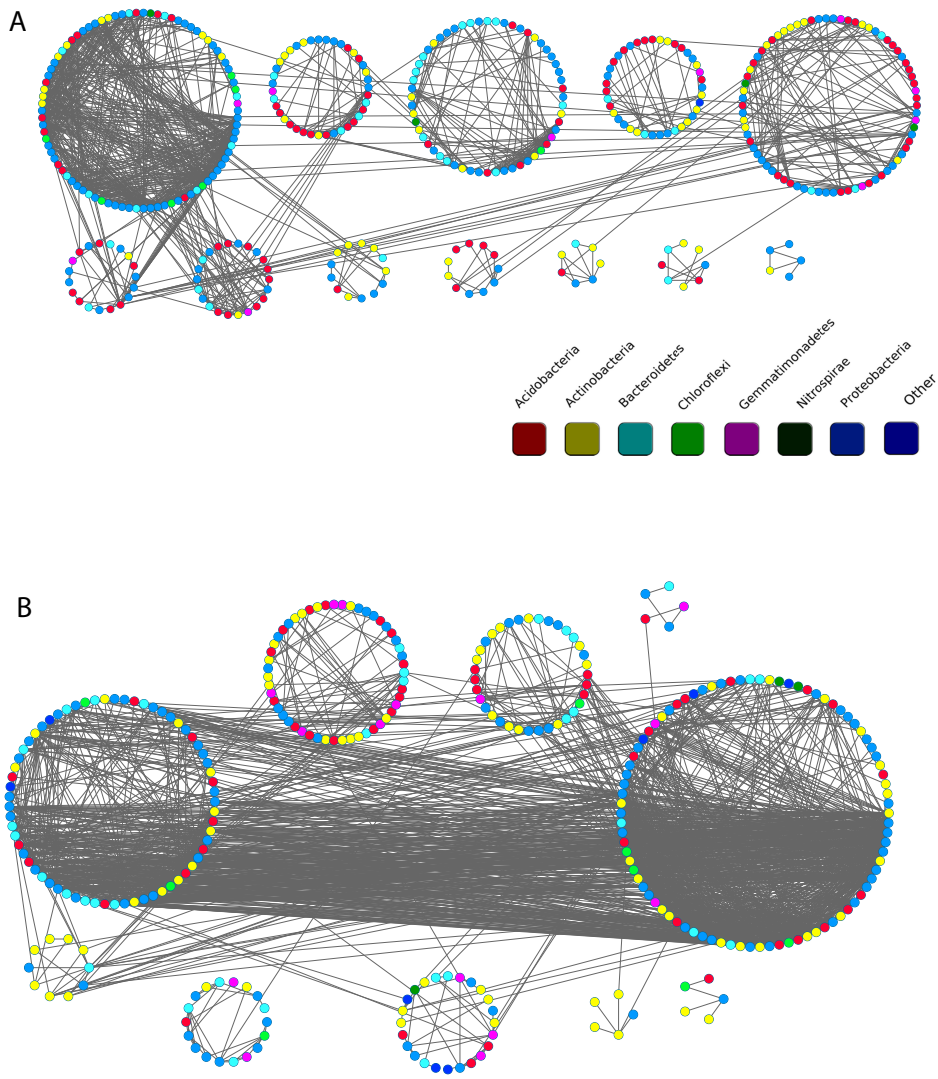


Figure 2. Ordination plots produced by principal coordinate analyses on Bray-Curtis distance matrices. Shown are the results of four PCoA each based on OTU data acquired by different clustering method. A, B, C, D correspond to Mothur, CROP, UCLUST and Swarm clustering methods, respectively. The samples belonging to same test site (Holvandi, Porijõgi, Sirgala, Songa) are surrounded by continuous line.

The networks acquired by MENAP are presented in **Figure 3**. It is to be remembered that modules sized less than 5 OTU-s are removed from the figure and any further analyses. The network acquired by using MENAP on Swarm clustered OTU abundance table stands out among others by having all modules moderately small and similar in sizes with sparse interconnections – while in figures depicting the results for analyses after other clustering methods there exists always at least a few modules that compared to others are large in size and have many interconnections with other modules.

The sparser network observed in the case of Swarm analysis might at least partially be explained by the fact that Swarm generated OTU-s that were quite often present only in one or few samples. Therefore the step in MENAP where we removed OTU-s that were not present in at least 5 samples had the most impact on Swarm analysis. The topological parameters of acquired networks are shown in **Table 5.**



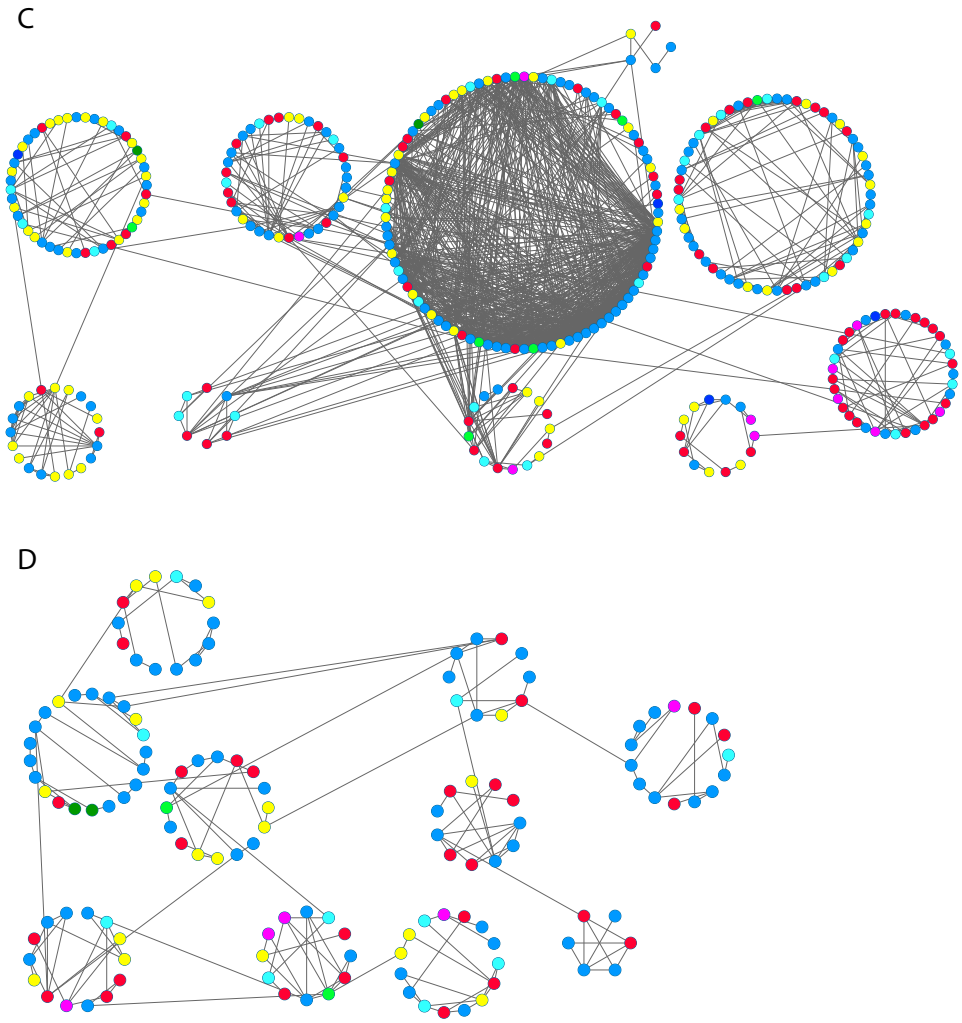


Figure 3. Plots of networks constructed using Molecular Ecological Network Analyses Pipeline on the OTU-data that was acquired for same samples by different OTU clustering methods. The networks are labeled such that A, B, C, D correspond to Mothur, CROP, UCLUST and Swarm clustering methods, respectively. The color legend available at 3A applies to every network.

Table 5. Topological properties of the empirical phylogenetic molecular ecological networks of microbial communities and their associated random phylogenetic molecular ecological networks. Averages calculated by 100 times randomly rewiring all of the links of the phylogenetic molecular ecological networks.

	Original OTU-s	Nodes	Edges	Average connectivity	Average geodesic distance	Average clustering coefficient	Modularity	Modules containing ≥ 5	R2 of powerlaw
Mothur	852	447	1143	0.72	7.47	0.28	0.65	12	0.91
CROP	670	387	1731	0.63	5.74	0.3	0.33	10	0.81
UCLUST	887	452	1267	0.59	8.76	0.26	0.54	10	0.80
Swarm	175	146	182	0.84	6.67	0.14	0.78	10	0.70

5.3. Differences in relationships between bacterial community structures and site-specific characteristics

Distance-based regression analysis was applied to identify soil variables that explain significant amounts of variation in bacterial community structure. For all sets of analyses pH and soil water content was related to bacterial community composition as shown in **Table 6.**

Table 6. Soil variables that explain significant amounts of variation in bacterial community structure. Distance based linear regression analyses results on Bray-Curtis distance matrices using soil chemical parameters as explanatory variables. Variables inserted into the forward selection using DISTLM analysis were water content (%), $pH_{KCl,}$ total nitrogen (%), phosphorus (mg/Kg, ammonium lactate extractable) and the amount of organic matter (based on LOI – loss on ignition). The table is formatted as following – following each clustering method the values included in final model are depicted with the percentage of variation explained levels for each variable included after all model terms a p-value level for the whole model is depicted.

Method	Soil variables	Variation explained (%)
Mothur	pH 22.9%, Dry weight 8.2% (<0.001)	31.1
CROP	pH 25.9%, Dry weight 8.8% (<0.001)	34.7
UCLUST	pH 23%, Dry weight 8.3% (<0.001)	31.3
Swarm	pH 10.7%, Dry weight 6.2% (<0.001)	16.9

When soil elemental concentrations were used in stepwise regression analysis as explanatory variables, the concentration values of boron and cadmium were included in the model for all sets of analyses, and with the exception of Swarm based sets also aluminium and iron. Swarm was also only set which had a model including copper as show in **Table 7.**

Table 7. Metallic element concentrations that explain significant amounts of variation in bacterial community structure. DISTLM analyses**.** Variables used were elemental concentrations of Al, B, Ca, Cd, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb and Zn.

Method	Metals	Variation explained (%)
Mothur	B 14.1%, Cd 13.2%, Al 7.4%, Fe 6.3% (<0.001)	40.9
CROP	B 15.3%, Cd 14%, Al 6.9%, Fe 7% (<0.001)	43.4
UCLUST	B 14.1%, Cd 13.1%, Al 7.7%, Fe 6% (<0.001)	41.1
Swarm	B 8.2%, Cd 7.9%, Cu 6.6% (<0.001)	22.7

5.4. Beta diversity results according to OTU clustering independent analysis

The term beta diversity has been used to refer to a wide variety of phenomena. Although all of these encompass some kind of compositional heterogeneity between places, many are not related to each other in any predictable way (Tuomisto 2010).

When looking at PCoA plots produced on the Kantorovich-Rubenstein distance matrix between the samples from all three publications (**Figure 4)** – it becomes visible how different the samples from Publication I are compared to the others. Reasons for these samples clustering together far from others is most likely caused by the different amplicons and/or sequencing platforms used in Publication I as compared to others.

It can be seen that although all the soil samples of Publication II were collected inside Estonia the same as Publication I on this PCoA they tend to be position much closer to Publication III samples, which were taken from a much wider set of geographical locations: including also Estonia, but also UK, Finland and Russia.

The description of V2 and partly V3 hyper-variable region of the 16S rRNA used in Publication I and its amplification and further sequencing on 454 platform is already provided in the Materials and methods section of this article. In Publications II and III the L-V6 and R-V6 primers (Gloor et al. 2010) were used to amplify the bacteria-specific V6 hyper variable region of the 16S rRNA. The sequencing in these cases was done using Illumina HiSeq 2000.

Figure 5 shows a PCoA plots where the soil samples used come only from the Publications II and III). This plot indicates that in the case of the same amplicon the OTU clustering irrelevant analysis is a viable way to compare samples. In case of **Figure 6** we can even observe similarities to the groupings produced in the case of OTU based strategies (**Figure 2)**.

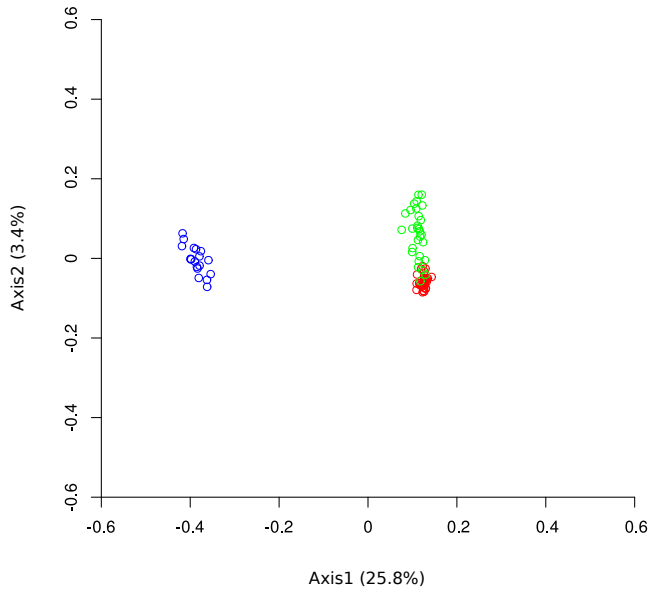


Figure 4. Ordination plot of PCoA based on Kantorovich-Rubenstein matrix. PCoA was produced from Kantorovich-Rubenstein distances calculated from phylogenetic placements. Placements were produced on samples from three datasets – green points denote samples from (Preem et al., 2012), red points samples from (Truu et al., 2017) and green points samples from (Ostonen et al., 2017)

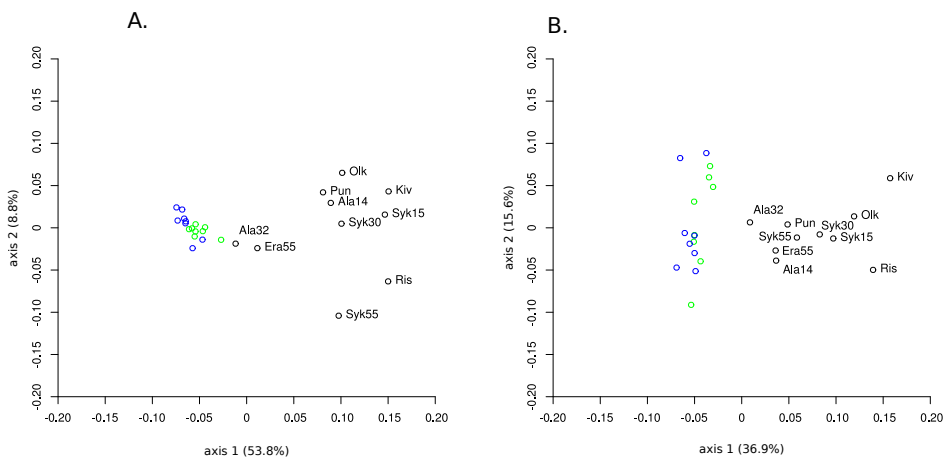


Figure 5. Ordination plots of PCoA based on Kantorovich-Rubenstein distance matrices. PCoA was produced from Kantorovich-Rubenstein distances calculated from phylogenetic placements. Placements were produced on samples from several datasets – colored points denote samples from (Truu et al., 2017) uncolored points with labels denote samples from (Ostonen et al., 2017). Plot A contains PCoA that is produced from bulk soil samples, and plot B a PCOA produced from rhizosphere samples. Green points denote samples from control plots and blue ones humidified plots

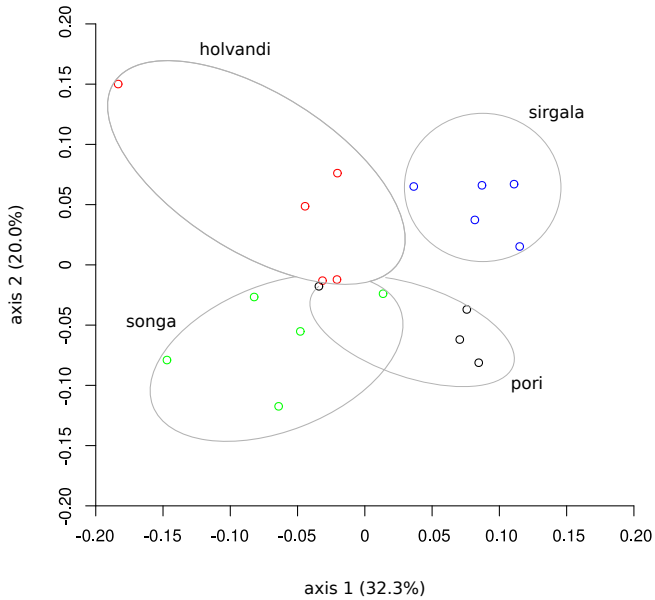


Figure 6. Ordination plot of PCoA based on Kantorovich-Rubenstein distance matrices. PCoA was produced from Kantorovich-Rubenstein distances calculated from phylogenetic placements. Placements were produced on samples from Publication I. This provides an alternative to OTU based analyses as depicted in Figure 2.

5.5. Comparison of clustering methods

While the composition and classification of OTU-s acquired by different OUT clustering methods can vary quite a lot the Kruskal-Wallis tests show that OTU numbers do not differ significantly between sites in any analysis, and most of the ecological conclusions acquired on the OTU datasets remain quite similar.

The estimation of overall similarity between bacterial communities seems to be similar among the different OTU clustering methods as proven by high congruency of Bray-Curtis distance matrices in CADM a test that has shown to have good performance for comparisons like this (Campbell et al. 2011). This similarity is also illustrated neatly in **Figure 2.** Showing that the distances between samples based on the OTU data remain similar in spite of different methods. The fact that that samples taken from the same region cluster together serves as a positive sanity check.

According to the Mantel test on distance matrices the most distant matrix from the others is the SWARM based one. The Swarm based analysis tends to also produce the most different biological conclusions compared to other clustering methods. So the SWARM algorithms ability to represent correct ecological/biological foundation compared to other methods comes to question.

It should be then taken into account that for processing OTU data gathered from current OTU clustering methods more robust statistical methods and analyses are preferable. For instance, PCoA, CADM and Kruskal-Wallis tests of OTU numbers and IS show similar output in case of different clustering methods used whereas analysis such as MENAP produces substantially different results.

The results of application of different clustering methods for the analysis of mock community data casts the clustering methods in worse light. Considering that the dataset of unique sequences was amplified/replicated by adding minor deviations (sequencing errors according to Balzer et al. (2011) model) from unique sequences so that the proportions of sequences derived from originals follow a pre-set distribution of copy numbers – the clustering method should cluster the sequences originating from original dataset into OTU-s following similar distribution.

The distributions of OTU-s garnered by clustering do not resemble the distribution in the mock set which sets their validity to doubt. The other possibility is the need for a more refined mock community generation method so that the process of adding deviations to the original seed sequence would follow more closely evolutional models or to create in addition to in silico mock communities mock communities – though that would be a resource intensive endeavor.

There is still some similarities in the OTU clustering methods in that all methods (with the exception of Swarm analysis where distance value was set to 4 nucleotides) have the same representative sequence for the largest OTU. Though this is not the sequence that is amplified the most in actual mock community.

PCoA plots acquired from OTU independent analysis behave in readily explainable manner and can be said to pass some sanity checks: we can determine from **Figure 5** how samples collected at Alatskivi and Erastvere sites (ala32 and era55) in Estonia (samples from (Publication III) group closer to the blue-green marked samples that are all collected in FAHM facility (Publication II) in Estonia compared to other named sites from more distant geographical locations.

In Figure 5 the samples from Publication I cluster farther from the samples from other publications. This can be expected as it uses different 16S rDNA region and sequencing platform compared to others. Taking this on account, the look at the **Figure 6** which is a PCoA on samples from Publication I similar to the PCoA plots in **Figure 2,** shows samples grouping into groups by site in an overall similar manner to OTU based analyses, although OTU based methods have a stronger separation of sites. The overall appearance of the plots is rather similar though the OTU independent analysis is still most dissimilar from others just after the SWARM based method.

The overall similarities in OTU based analysis and OTU-independent analysis support each-other

In case of different OTU clustering methods the Swarm analysis stands out as most different from others also could be said to be most un-trustworthy. The

basic assumption of Swarm methodology gives for an elegant basis for a computationally efficient clustering algorithm but should be investigated more. It seems instinctually make sense that evolutionary processes proceed stepwise to form a cluster of related/similar progeny centered around an ancestor – but how well can we assume that the ancestors and progenies are to our comfort represented in environmental samples in a neat enough chains with no lacking intermediates causing major gaps. The exact nature of mutations that cause the generation of new sequences is something that would also affect the algorithm especially the selection of suitable k parameter. The troubles of choosing adequate parameter k for programs run are demonstrated in the mock community analysis.

Mothur based analysis although not providing as different results as others but the simple average-neighbor clustering algorithm based on distance matrix and the generation of such distance matrix would be prohibitively computationally expensive for larger datasets. As of such UCLUST and CROP stand out as clustering methods of choice.

The similarity of PCoA plots alongside to the fact that they look as expected lends also support to this as a highly robust one (the suitability for ordination methods for different tasks has been disputed and discussed such as (Minchin 1987; Ruokolainen and Salo 2006) but in our case PCoA produces verifiably "sane" results). The thesis demonstrates the benefits of robust methods such as this – which are able to work consistently even with some differences that are produced by different clustering methods. On the other hand, network analysis results obtained with MENAP shows itself as highly sensitive to such differences and its results therefore are suspect.

6. CONCLUSIONS

Different clustering methods will mostly reach the same ecological conclusions for the same dataset. The biggest differences generated from others come from SWARM method.

OTU clustering independent analysis results being similar to those acquired from clustering based strengthens the idea that all these methods are capable of accessing ecologically meaningful patterns in bacterial communities.

As a recommendation, UCLUST and CROP methods stand out while SWARM which produces most different results from the consensus elicits some doubt. Mothur clustering method which produces comparable results to others, will experience quickly growing needs for computational resources for larger datasets so its usability will be more limited in this regard.

MENAP networks show up radically different for each OUT clustering method – this seems to show more of an issue with the MENAP methodology being overly sensitive. The thesis shows the practicality of using robust data analysis methods when dealing with OTU data acquired by clustering.

All OTU clustering methods fail to satisfactorily reproduce the build of mock community – it might stand that this part of analysis should require a revisit – a experiment with better mock community generation protocol and more complex mock community.

REFERENCES

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004) Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. J Bacteriol 186:2629–2635. doi: 10.1128/JB.186.9.2629-2635.2004

Anderson M (2001) A new method for nonparametric multivariate analysis of variance. Austral Ecol 26:32–46.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res 40:e94–e94. doi: 10.1093/nar/gks251

Baldrian P (2016) Forest microbiome: diversity, complexity and dynamics. FEMS Microbiol Rev fuw040. doi: 10.1093/gbe/evw245

Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M, Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2004) Pfam: The protein families database. Nucleic Acids Res 42:138–141. doi: 10.1093/nar/gkt1223

Bodilis J, Nsigue-Meilo S, Besaury L, Quillet L (2012) Variable copy number, intragenomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. PLoS One. doi: 10.1371/journal.pone.0035647

Cai Y, Sun Y (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Res 39:e95. doi: 10.1093/nar/gkr349

Campbell V, Legendre P, Lapointe F-J (2011) The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. BMC Evol Biol 11:64. doi: 10.1186/1471-2148-11-64

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley G a, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone C a, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters W a, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–6. doi: 10.1038/nmeth.f.303

Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H (2013) A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. PLoS One 8:e70837. doi: 10.1371/journal.pone.0070837

Cheng L, Walker AW, Corander J (2012) Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Res 40:1–10. doi: 10.1093/nar/gks227

Chodak M, Gołębiewski M, Morawska-Płoskonka J, Kuduk K, Niklińska M (2015) Soil chemical properties affect the reaction of forest soil bacteria to drought and rewetting stress. Ann Microbiol 65:1627–1637. doi: 10.1007/s13213-014-1002-0

Colin Y, Nicolitch O, Turpault MP, Uroz S (2017) Mineral types and tree species determine the functional and taxonomic structures of forest soil bacterial communities. Appl Environ Microbiol. doi: 10.1128/AEM.02684-16

Cong J, Yang Y, Liu X, Lu H, Liu X, Zhou J, Li D, Yin H, Ding J, Zhang Y (2015) Analyses of soil microbial community compositions and functional genes reveal potential consequences of natural forest succession. Sci Rep 5:10007. doi: 10.1038/srep10007

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 13:696–710. doi: 10.1093/bib/bbs070

Deng Y, Jiang Y, Yang Y, He Z, Luo F, Zhou J (2012) Molecular ecological network analyses. BMC Bioinformatics 13:113. doi: 10.1186/1471-2105-13-113

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–72. doi: 10.1128/AEM.03006-05

Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G (2013) High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods 95:401–414. doi: 10.1016/j.mimet.2013.08.011

Edgar RC (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucleic Acids Res 32:380–385. doi: 10.1093/nar/gkh180

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200. doi: 10.1093/bioinformatics/btr381

Edwards U, Rogall T, Blöcker H, Emde M, Böttger EC (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res 17:7843–7853. doi: 10.1093/nar/17.19.7843

Escobar-Zepeda A, De León AVP, Sanchez-Flores A (2015) The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Front Genet 6:1–15. doi: 10.3389/fgene.2015.00348

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Fraley C, Raftery a E (1998) How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Comput J 41:578–588. doi: 10.1093/comjnl/41.8.578

Fraley C, Raftery a E (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631. doi: 10.1198/016214502760047131

Giles M, Morley N, Baggs EM, Daniell TJ (2012) Soil nitrate reducing processes – Drivers, mechanisms for spatial variation, and significance for nitrous oxide production. Front Microbiol 3:1–16. doi: 10.3389/fmicb.2012.00407

Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R, Reid G (2010) Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. PLoS One 5:e15406. doi: 10.1371/journal.pone.0015406

Handelsman J, Rondon M, Brady S, Clardy J (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27:611–8. doi: 10.1093/bioinformatics/btq725

Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: Predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37:101–105. doi: 10.1093/nar/gkp327

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12:1889–98. doi: 10.1111/j.1462-2920.2010.02193.x

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. doi: 10.1186/1471-2105-11-119

Jeanbille M, Buée M, Bach C, Cébron A, Frey-Klett P, Turpault MP, Uroz S (2016) Soil Parameters Drive the Structure, Diversity and Metabolic Potentials of the Bacterial Communities Across Temperate Beech Forest Soil Sequences. Microb Ecol 71:482–493. doi: 10.1007/s00248-015-0669-5

Jonsson V, Österlund T, Nerman O, Kristiansson E (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics 17:78. doi: 10.1186/s12864-016-2386-y

Kanehisa M, Goto S, Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. doi: 10.1093/nar/27.1.29

Kim D, Hahn AS, Wu SJ, Hanson NW, Konwar KM, Hallam SJ (2015) FragGeneScan-plus for scalable high-throughput short-read open reading frame prediction. 2015 IEEE Conf Comput Intell Bioinforma Comput Biol CIBCB 2015 1–8. doi: 10.1109/CIBCB.2015.7300341

Koeppel AF, Wu M (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Res 41:5175–5188. doi: 10.1093/nar/gkt241

Kowalchuk G a, Speksnijder AGCL, Zhang K, Goodman RM, van Veen J a (2007) Finding the needles in the metagenome haystack. Microb Ecol 53:475–85. doi: 10.1007/s00248-006-9201-2

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12:118–123. doi: 10.1111/j.1462-2920.2009.02051.x

Legendre P, Laponte F-J, Lapointe F-J (2004) Assessing congruence among distance matrices: single-malt scotch whiskies revisited. Aust N Z J Stat 46:615–629. doi: 10.1111/j.1467-842X.2004.00357.x

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. doi: 10.1093/bioinformatics/btl158

Lladó SF, López-Mondéjar R, Baldrian P (2017) Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. 81:1–27. doi: 10.1128/mmbr.00063-16

López-Lozano NE, Heidelberg KB, Nelson WC, García-Oliva F, Eguiarte LE, Souza V (2013) Microbial secondary succession in soil microcosms of a desert oasis in the Cuatro Cienegas Basin, Mexico. PeerJ 1:e47. doi: 10.7717/peerj.47

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. doi: 10.1186/s13059-014-0550-8

Lozupone C, Knight R (2005) UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol 71:8228–8235. doi: 10.1128/AEM.71.12.8228

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm : robust and fast clustering method for amplicon-based studies PrePrints PrePrints. PeerJ 1–12. doi: http://dx.doi.org/10.7287/peerj.preprints.386v1

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

Maidak B, Cole J, Lilburn T, Parker C, Saxman P, Farris R, Garrity G, Olsen G, Schmidt T, Tiedje J (2001) The RDP-II (ribosomal database project). Nucleic Acids Res. doi: 10.1093/nar/29.1.173

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) CDD: A Conserved Domain Database for protein classification. Nucleic Acids Res 33:192–196. doi: 10.1093/nar/gki069

Marco D (ed. . (2010) Metagenomics: Theory, Methods and Applications | Book. Caister Academic Press, Norwich

Margulies M, Egholm M, Altman WWE, Attiya S, Bader JS, Bemben L a, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt K a, Volkmer G a, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. doi: 10.1038/nature03959

Matsen F a, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538. doi: 10.1186/1471-2105-11-538

Matsen IV FA, Evans SN (2013) Edge Principal Components and Squash Clustering: Using the Special Structure of Phylogenetic Placement Data for Sample Comparison. PLoS One. doi: 10.1371/journal.pone.0056859

Matthew Stephens (2000) Bayesian Analysis of Mixture Models with an Unknown Number of Components- An Alternative to Reversible Jump Methods. Ann Stat 28:40–74.

McArdle B, Anderson M (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82:290–297.

Mendes R, Garbeva P, Raaijmakers JM (2013) The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiol Rev 37:634–663. doi: 10.1111/1574-6976.12028

Minchin PR (1987) An evaluation of the relative robustness of techniques for ecological ordination. 89–107.

MIRARAB S, NGUYEN N, WARNOW T SEPP: SATé-Enabled Phylogenetic Placement. Biocomput 2012 247–258. doi: 10.1142/9789814366496_0024

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, Von Mering C, Doerks T, Jensen LJ, Bork P (2009) eggNOG v2.0: Extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38:190–195. doi: 10.1093/nar/gkp951

Muyzer G, De Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol 59:695–700. doi: 0099-2240/93/030695-06$02.00/0

Nacke H, Goldmann K, Schöning I, Pfeiffer B, Kaiser K, Villamizar GAC, Schrumpf M, Buscot F, Daniel R, Wubet T (2016) Fine spatial scale variation of soil microbial communities under European beech and Norway spruce. Front Microbiol. doi: 10.3389/fmicb.2016.02067

Nalbantoglu U, Cakar A, Dogan H, Abaci N, Ustek D, Sayood K, Can H (2014) Metagenomic analysis of the microbial community in kefir grains. Food Microbiol 41:42–51. doi: 10.1016/j.fm.2014.01.014

Nazaries L, Murrell JC, Millard P, Baggs L, Singh BK (2013a) Methane, microbes and models: Fundamental understanding of the soil methane cycle for future predictions. Environ Microbiol 15:2395–2417. doi: 10.1111/1462-2920.12149

Nazaries L, Pan Y, Bodrossy L, Baggs EM, Millard P, Murrell JC, Singh BK (2013b) Evidence of microbial regulation of biogeochemical cycles from a study on methane flux and land use change. Appl Environ Microbiol 79:4031–4040. doi: 10.1128/AEM.00095-13

Nebel M, Pfabel C, Stock A, Dunthorn M, Stoeck T (2011) Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. Environ Microbiol Rep 3:154–8. doi: 10.1111/j.1758-2229.2010.00200.x

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15:387–96. doi: 10.1093/dnares/dsn027

Ostonen, Ivika; Truu, Marika; Helmisaari, Heljä-Sisko; Lukac, Martin; Vanguelova, Elena; Godbold, Douglas L; Lõhmus, Krista; Zang, Ulrich, Tedersoo, Leho; Preem, Jens-Konrad; Rosenvald, Katrin; Aosaar, Jürgen; Armolaitis, Kestutis; Frey, Jane; Kabral, Nai J (2017) Adaptive root foraging strategies along a boreal-temperate forest gradient.

Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rül;ckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691–5702. doi: 10.1093/nar/gki866

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res 35:e130. doi: 10.1093/nar/gkm760

Petrosino JF, Highlander S, Luna RA, Gibbs R a, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55:856–66.

Preem J-K, Truu J, Truu M, Mander Ü, Oopkaup K, Lõhmus K, Helmisaari H-S, Uri V, Zobel M (2012) Bacterial community structure and its relationship to soil physico-chemical characteristics in alder stands with different management histories. Ecol Eng 49:10–17. doi: 10.1016/j.ecoleng.2012.08.034

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–96. doi: 10.1093/nar/gkm864

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6:1–6. doi: 10.1038/nmeth.1361

Rho M, Tang H, Ye Y (2010) FragGeneScan: Predicting genes in short and error-prone reads. Nucleic Acids Res 38:1–12. doi: 10.1093/nar/gkq747

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components – Discussion. J R Stat Soc Ser B (Statistical Methodol 59:731–792. doi: 10.1111/1467-9868.00095

Ruokolainen L, Salo K (2006) Differences in performance of four ordination methods on a complex vegetation dataset. 269–275.

Saggar S, Jha N, Deslippe J, Bolan NS, Luo J, Giltrap DL, Kim DG, Zaman M, Tillman RW (2013) Denitrification and N2O: N2 production in temperate grasslands: Processes, measurements, modelling and mitigating negative impacts. Sci Total Environ 465:173–195. doi: 10.1016/j.scitotenv.2012.11.050

Schloss PD, Handelsman J (2006a) Toward a Census of Bacteria in Soil. PLoS Comput Biol 2:e92. doi: 10.1371/journal.pcbi.0020092

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71:1501–1506. doi: 10.1128/AEM.71.3.1501-1506.2005

Schloss PD, Handelsman J (2006b) Introducing TreeClimber , a Test To Compare Microbial Community Structures. Appl Environ Microbiol 72:2379–2384. doi: 10.1128/AEM.72.4.2379

Schloss PD, Larget BR, Handelsman J (2004) Integration of microbial ecology and statistics: A test to compare gene libraries. Appl Environ Microbiol 70:5485–5492. doi: 10.1128/AEM.70.9.5485-5492.2004

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski R a, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–41. doi: 10.1128/AEM.01541-09

Schouls LM, Schot CS, Jacobs JA (2003) Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group. J Bacteriol 185:7241–7246. doi: 10.1128/JB.185.24.7241

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O (2007) TIGRFAMs and Genome Properties: Tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35:D260–D264. doi: 10.1093/nar/gkl1043

Seo DC, DeLaune RD (2010) Fungal and bacterial mediated denitrification in wetlands: Influence of sediment redox condition. Water Res 44:2441–2450. doi: 10.1016/j.watres.2010.01.006

Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5:209. doi: 10.3389/fpls.2014.00209

Siles JA, Margesin R (2016) Abundance and Diversity of Bacterial, Archaeal, and Fungal Communities Along an Altitudinal Gradient in Alpine Forest Soils: What Are the Driving Factors? Microb Ecol 72:207–220. doi: 10.1007/s00248-016-0748-2

Singleton DR, Furlong M a., Rathbun SL, Whitman WB (2001) Quantitative Comparisons of 16S rRNA Gene Sequence Libraries from Environmental Samples. Appl Environ Microbiol 67:4374–4376. doi: 10.1128/AEM.67.9.4374-4376.2001

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci 103:12115–12120. doi: 10.1073/pnas.0605127103

Soueidan H, Nikolski M (2015) Machine learning for metagenomics: methods and tools.

Stackebrandt E, Goebel BM (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol 44:846–849. doi: 10.1099/00207713-44-4-846

Steven B, Gallegos-Graves LV, Starkenburg SR, Chain PS, Kuske CR (2012) Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. Environ Microbiol Rep 4:248–256. doi: 10.1111/j.1758-2229.2012.00328.x

Steven N. Evans, Matsen FA, Evans SN, Matsen FA (2012) The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. J R Stat Soc Ser B (Statistical Methodol 74:569–592. doi: 10.1111/j.1467-9868.2011.01018.x

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, Mai V (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Brief Bioinform 13:107–21. doi: 10.1093/bib/bbr009

Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Res 37:e76. doi: 10.1093/nar/gkp285

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov A V, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. doi: 10.1186/1471-2105-4-41

Tonge DP, Pashley CH, Gant TW (2014) Amplicon -based metagenomic analysis of mixed fungal samples using proton release amplicon sequencing. PLoS One. doi: 10.1371/journal.pone.0093849

Truu M, Ostonen I, Preem J-K, Lõhmus K, Nõlvak H, Ligi T, Rosenvald K, Parts K, Kupper P, Truu J (2017) Elevated Air Humidity Changes Soil Bacterial Community Structure in the Silver Birch Stand. Front Microbiol 8:1–15. doi: 10.3389/fmicb.2017.00557

Tuomisto H (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography (Cop) 33:2–22. doi: 10.1111/j.1600-0587.2009.05880.x

Turner TR, James EK, Poole PS, Gilbert J, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Al. E (2013) The plant microbiome. Genome Biol 14:209. doi: 10.1186/gb-2013-14-6-209

Uroz S, Buée M, Deveau A, Mieszkin S, Martin F (2016a) Ecology of the forest microbiome: Highlights of temperate and boreal ecosystems. Soil Biol Biochem 103:471–488. doi: 10.1016/j.soilbio.2016.09.006

Uroz S, Oger P, Tisserand E, Cébron A, Turpault M-P, Buée M, De Boer W, Leveau JHJ, Frey-Klett P (2016b) Specific impacts of beech and Norway spruce on the structure and diversity of the rhizosphere and soil microbial communities. Sci Rep 6:27756. doi: 10.1038/srep27756

Yeung KY, Fraley C, Murua a., Raftery a. E, Ruzzo WL (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17:977–987. doi: 10.1093/bioinformatics/17.10.977

Zhou J, Deng Y, Shen L, Wen C, Yan Q, Ning D, Qin Y, Xue K, Wu L, He Z, Voordeckers JW, Nostrand JD Van, Buzzard V, Michaletz ST, Enquist BJ, Weiser MD, Kaspari M, Waide R, Yang Y, Brown JH (2016) Temperature mediates continental-scale diversity of microbes in forest soils. Nat Commun 7:12083. doi: 10.1038/ncomms12083

SUMMARY IN ESTONIAN

Amplikoni põhine metsamuldade bakterikoosluse analüüs

Muld, olles paljude ökoloogiliste protsesside keskmes, on arusaadavalt pärinud tähelepanu ja uurimist mitmest vallast. Muldade rikkalike mikrobioomide uurimist on siiani takistanud tõsiasi, et enamik mulla mikroobe on kultiveerimatud. Seda kitsaskohta aitab leevendada metagenoomika, mis tähistab uurimistööd otse keskkonnaproovidest eraldatud geneetilise materjaliga. Selline lähenemine võimaldab uurida ka kultiveerimatuid mikroobe.

Publikatsioonid, millel käesolev töö põhineb, käsitlesid peamiselt metsamuldade mikrobioome ning lisaks mõningal määral ka taimede mikrobioome (täpsemalt risosfääri kooslusi). Publikatsioonides kasutati 16S rDNA amplikoni põhiseid metagenoomilisi meetodeid mulla mikrobioomi kirjeldamiseks.

Selliste andmete kasutamiseks on levinud meetodid, mille abil grupeeritakse (klasterdatakse) kogutud DNA järjestused *ad-hoc* taksonoomilistesse üksustesse nn. OTU-desse (*Operational Taxonomic Unit*). Nii võib OTU-desse klasterdatud järjestusi kasutades hinnata bakterikoosluste mitmekesisust ja liigilist koostist. Saadud OTU-de arvukuse numbreid saab kasutada mitmesugustes erinevates analüüsides kui asendajaid konventsionaalsematele taksonoomilistele üksustele. Niisama kiire, kui on olnud uute sekveneerimistehnoloogiate areng, on ka olnud uute tööriistade arvu kasv – viimase kümnendi jooksul on loodud hulk programme, mis on mõeldud eelpoolmainitud OTU-de moodustamiseks DNA järjestuste andmetest.

Käesolev töö keskendub sellele, kuidas mõjutavad erinevad OTU loomise meetodid edasist andmeanalüüsi ning lõplikke järeldusi. Selleks kasutati järjestusandmeid artiklist "Bacterial community structure and its relationship to soil physico-chemical characteristics in alder stands with different management histories" (originaalartikkel I) ning erinevaid OTU klasterdamise meetodeid.

OTU-d loodi Mothuri, UCLUST, CROP ja Swarm algoritmide abil – seejärel viidi läbi koosluste erinevad statistilised analüüsid. Paremaks OTU klasterdamismeetodite võrdluseks loodi ka *in silico* bakterikooslus, mida käsitleti samal viisil kui artikkel I järjestusandmeid. Võrdluseks viidi läbi OTU klasterdamisest sõltumatuks analüüs, mis kujutas endast mitmemõõtmelist analüüsi järjestuste fülogeneetilistest paigutustest arvutatud Kantorovich-Rubenstein kauguste põhjal. Sellesse analüüsi oli täiendavateks võrdlusteks lisaks kaasatud ka originaalartiklite II ja III andmed.

Erinevate meetoditega saadud OTU andmete analüüs andis üldjoontes samasuguseid tulemusi. Seda visualiseerivad hästi töös olevad mitmemõõtmelise analüüsi joonised. Kuigi proovikohtade OTUde arvud ja mitmekesisusindeksid olid meetodite vahel erinevad, siis statistilised testid iga meetodi siseselt viisid samale järelduse, et proovikohad üksteiset ei erine OTU arvu ja mitmekesisusindeksite poolest. Samuti olid sarnased regressioonanalüüside tulemused. Teis-

test erinevaimad olid SWARM meetodiga läbi viidud analüüside tulemused. Eri meetodite abil saadud OTU-de alusel loodud Bray-Curtis kaugusmaatriksid olid CADM analüüsi alusel tugevalt kongruentsed. Kõige erinevam maatriks oli Mantel testi alusel jällegi SWARM meetodi tulem.

OTU klasterdamise vaba analüüsi tulemused kinnitasid eri klasterdamismeetoditega saadud analüüsi tulemusi, kuid *in silico* koosluse puhul koosluses esialgselt olnud sünteetiliste liikide sagedusi ei suutnud korrektselt taastada ükski klasterdamismeetod – selle taga võivad olla ka probleemid *in silico* koosluse valmistamisel.

Arvestades, et kõige rohkem erinevusi andmeanalüüsi tulemustes kaasnes SWARM meetodi kasutamisega ning Mothuri kaugusmaatrikseid kasutav klasterdamismeetod on suuremate andmehulkade kasutamisel piiratud, jäävad antud töö tulemusena soovitusena sõelale CROP ja UCLUST meetodid.

Lisaks peab ka mainima, et kui eri klasterdamismeetoditega loodud OTU tabelite analüüsid andsid samasuguseid tulemusi OTU numbrite, OTU alusel loodud kaugusmaatriksitele tehtud ordinatsioonidele ning neile tehtud lineaarsetele regressioonidele keskkonnanäitajatele, siis MENAP võrgustikuanalüüs andis eri meetodite puhul drastiliselt erinevaid tulemusi – see näitab robustsemate meetodite kasulikkust OTU andmete analüüsil.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Education and Science of Estonia (grant SF0180127s08), the Estonian Research Council (grant IUT2-16); and the EU through the European Regional Development Fund (Centre of Excellence ENVIRON and Centre of Excellence EcolChange).

I want to thank my supervisors Prof. Jaak Truu and Prof. Ülo Mander for their help and guidance.

I am also very thankful to my colleagues who have offered me their advice, insight and company.

Finally I would like to thank my wife who has encouraged and supported me during my studies.

PUBLICATIONS

CURRICULUM VITAE

Name: Jens-Konrad Preem
Date of birth: March 14, 1984
Citizenship: Estonian
Phone: +372 554 3263
E-mail: jpreem@ut.ee

Education:

2006–2009	University of Tartu – MSc in Biology
2002–2006	University of Tartu – Bachelor in Biology

Occupation:

since 2016	Universtity of Tartu, Faculty of Science and Technology, Institute of Ecology and Earth Sciences, specialist
2010–2016	Competence Centre on Health Technologies, researcher
2014–2015	Universtity of Tartu, Faculty of Science and Technology, Institute of Ecology and Earth Sciences, junior researcher
2013–2014	Universtity of Tartu, Faculty of Science and Technology, Institute of Ecology and Earth Sciences, specialist

Memberships: Estonian Society for Microbiology

Publications used in thesis:

J.-K. Preem, J. Truu, M. Truu, Ü. Mander, K. Oopkaup, K. Lõhmus, H.-S. Helmisaari, V. Uri, and M. Zobel, "Bacterial community structure and its relationship to soil physico-chemical characteristics in alder stands with different management histories," Ecol. Eng., vol. 49, no. null, pp. 10–17, Dec. 2012.

M. Truu, I. Ostonen, **J.-K. Preem**, K. Lõhmus, H. Nõlvak, T. Ligi, K. Rosenvald, K. Parts, P. Kupper, and J. Truu, "Elevated Air Humidity Changes Soil Bacterial Community Structure in the Silver Birch Stand," Front. Microbiol., vol. 8, no. April, pp. 1–15, 2017.

I. Ostonen, M. Truu, M. Lukac, W. Borken, E. Vanguelova, D. L. Godbold, L. Krista, U. Zang, L. Tedersoo, **J.-K. Preem**, and K. Rosenvald, "Adaptive root foraging strategies along a boreal – temperate forest gradient," New Phytol., vol. 215, pp. 977–991, 2017.

Other publications:

R. Mändar, M. Punab, N. Borovkova, E. Lapp, R. Kiiker, P. Korrovits, A. Metspalu, K. Krjutškov, H. Nõlvak, **J.-K. Preem**, K. Oopkaup, A. Salumets, and J. Truu, "Complementary seminovaginal microbiome in couples," Res. Microbiol., vol. 166, no. 5, pp. 440–7, 2015.

R. Mändar, M. Punab, P. Korrovits, S. Türk, K. Ausmees, E. Lapp, **J.-K. Preem**, K. Oopkaup, A. Salumets, and J. Truu, “Seminal microbiome in men with and without prostatitis,” Int. J. Urol., vol. 24, no. 3, pp. 211–216, Mar. 2017.

K. Tiirik, H. Nõlvak, K. Oopkaup, M. Truu, **J.-K. Preem**, A. Heinaru, and J. Truu, “Characterization of the bacterioplankton community and its antibiotic resistance genes in the Baltic Sea,” Biotechnol. Appl. Biochem., vol. 61, no. 1, pp. 23–32, 2014.

V. Vengerfeldt, K. Špilka, M. Saag, **J.-K. Preem**, K. Oopkaup, J. Truu, and R. Mändar, “Highly Diverse Microbiota in Dental Root Canals in Cases of Apical Periodontitis (Data of Illumina Sequencing),” J. Endod., vol. 40, no. 11, pp. 1778–1783, 2014.

ELULOOKIRJELDUS

Nimi: Jens-Konrad Preem
Sünniaeg: 14. märts 1984
Kodakondsus: Eesti
Telefon: +372 554 3263
E-Post: jpreem@ut.ee

Haridus:

2006–2009 Magistrikraad bioloogias, Tartu Ülikool
2002–2006 Bakalaureusekraad bioloogias, Tartu Ülikool

Teenistuskäik:

alates 2016 Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Ökoloogia ja maateaduste instituut, keskkonnatehnoloogia spetsialist
2010–2016 Tervisetehnoloogiate Arenduskeskus AS, teadur
2014–2015 Tartu Ülikool, Loodus- ja tehnoloogiateaduskond, Tartu Ülikooli Ökoloogia- ja Maateaduste Instituut, keskkonnatehnoloogia nooremteadur
2013–2014 Tartu Ülikool, Loodus- ja tehnoloogiateaduskond, Tartu Ülikooli Ökoloogia- ja Maateaduste Instituut, keskkonnatehnoloogia spetsialist

Teadusorganisatsioonid: Eesti Mikrobioloogide Ühenduse liige

Väitekirjaga seotud publikatsioonid:

J.-K. Preem, J. Truu, M. Truu, Ü. Mander, K. Oopkaup, K. Lõhmus, H.-S. Helmisaari, V. Uri, and M. Zobel, "Bacterial community structure and its relationship to soil physico-chemical characteristics in alder stands with different management histories," *Ecol. Eng.*, vol. 49, no. null, pp. 10–17, Dec. 2012.

M. Truu, I. Ostonen, **J.-K. Preem**, K. Lõhmus, H. Nõlvak, T. Ligi, K. Rosenvald, K. Parts, P. Kupper, and J. Truu, "Elevated Air Humidity Changes Soil Bacterial Community Structure in the Silver Birch Stand," *Front. Microbiol.*, vol. 8, no. April, pp. 1–15, 2017.

I. Ostonen, M. Truu, M. Lukac, W. Borken, E. Vanguelova, D. L. Godbold, L. Krista, U. Zang, L. Tedersoo, **J.-K. Preem**, and K. Rosenvald, "Adaptive root foraging strategies along a boreal – temperate forest gradient," *New Phytol.*, vol. 215, pp. 977–991, 2017.

Muud publikatsioonid:

R. Mändar, M. Punab, N. Borovkova, E. Lapp, R. Kiiker, P. Korrovits, A. Metspalu, K. Krjutškov, H. Nõlvak, **J.-K. Preem,** K. Oopkaup, A. Salumets, and J. Truu, "Complementary seminovaginal microbiome in couples," *Res. Microbiol.*, vol. 166, no. 5, pp. 440–7, 2015.

R. Mändar, M. Punab, P. Korrovits, S. Türk, K. Ausmees, E. Lapp, **J.-K. Preem**, K. Oopkaup, A. Salumets, and J. Truu, “Seminal microbiome in men with and without prostatitis,” *Int. J. Urol.*, vol. 24, no. 3, pp. 211–216, Mar. 2017.

K. Tiirik, H. Nõlvak, K. Oopkaup, M. Truu, **J.-K. Preem**, A. Heinaru, and J. Truu, “Characterization of the bacterioplankton community and its antibiotic resistance genes in the Baltic Sea,” *Biotechnol. Appl. Biochem.*, vol. 61, no. 1, pp. 23–32, 2014.

V. Vengerfeldt, K. Špilka, M. Saag, **J.-K. Preem**, K. Oopkaup, J. Truu, and R. Mändar, “Highly Diverse Microbiota in Dental Root Canals in Cases of Apical Periodontitis (Data of Illumina Sequencing),” *J. Endod.*, vol. 40, no. 11, pp. 1778–1783, 2014.

DISSERTATIONES TECHNOLOGIAE CIRCUMIECTORUM UNIVERSITATIS TARTUENSIS

1. **Sille Teiter.** Emission rates of N_2O, N_2, CH_4 and CO_2 in riparian grey alder forests and subsurface flow constructed wetlands. Tartu, 2005, 134 p.
2. **Kaspar Nurk.** Relationships between microbial characteristics and environmental conditions in a horizontal subsurface flow constructed wetland for wastewater treatment. Tartu, 2005, 123 p.
3. **Märt Öövel.** Performance of wastewater treatment wetlands in Estonia. Tartu, 2006, 148 p.
   Sergei Yurchenko. Determination of some carcinogenic contaminants in food. Tartu, 2006, 143 p. Published in *Dissertation Chimicae Universitatis Tartuensis,* 51.
4. **Alar Noorvee.** The applicability of hybrid subsurface flow constructed wetland systems with re-circulation for wastewater treatment in cold climates. Tartu, 2007, 117 p.
   Ülle Jõgar. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p. Published in *Dissertation Biologicae Universitatis Tartuensis*, 139.
5. **Christina Vohla.** Phosphorus removal by various filter materials in subsurface flow constructed wetlands. Tartu, 2008, 103 p.
6. **Martin Maddison.** Dynamics of phytomass production and nutrient standing stock of cattail and its use for environment-friendly construction. Tartu, 2008, 87 p.
7. **Marika Truu.** Impact of land use on microbial communities in Estonian soils. Tartu, 2008, 126 p.
8. **Elar Põldvere.** Removal of organic material, nitrogen and phosphorus from wastewater in hybrid subsurface flow constructed wetlands. Tartu, 2009, 107 p.
9. **Margit Kõiv.** Treatment of landfill leachate and municipal wastewater in subsurface flow filters using mineralized peat and hydrated oil shale ash. Tartu, 2010, 147 p.
10. **Jaanis Juhanson.** Impact of phytoremediation and bioaugmentation on the microbial community in oil shale chemical industry solid waste. Tartu, 2010, 95 p.
    Aare Selberg. Evaluation of environmental quality in Northern Estonia by the analysis of leachate. Tartu, 2010, 117 p. Published in *Dissertation Chimicae Universitatis Tartuensis,* 99.
11. **Riho Mõtlep.** Composition and diagenesis of oil shale industrial solid wastes. Tartu, 2010, 127 p.
12. **Igor Zaytsev.** Bioaugmentation in LWA-filled horizontal subsurface flow filters for wastewater treatment: Impact of flow regime, temperature and donor system Tartu, 2010, 97 p.

13. **Siiri Velling.** Microbial BOD biosensor for wastewater analysis. Tartu, 2011, 79 p.
14. **Riina Lepik.** Biodegradability of phenolic compoundsas single and mixed substrates by activated sludge. Tartu, 2011, 153 p.
15. **Liis Marmor.** Ecology and bioindicative value of epiphytic lichens in relation to air pollution and forest continuity. Tartu, 2011, 98 p.
16. **Martin Liira.** Active filtration of phosphorus in Ca-rich hydrated oil shale ash: precipitation mechanisms and recovery. Tartu, 2012, 84 p.
17. **Kristjan Karabelnik.** Advanced design and management of hybrid constructed wetlands: environmental and water purification effects. Tartu, 2012, 128 p.
18. **Hiie Nõlvak.** Influence of qPCR workflow on target gene enumeration from environmental samples in the case of bioremediation potential estimation. Tartu, 2012, 136 p.
19. **Merlin Raud.** Study of semi-specific BOD biosensors for biosensor-array. Tartu, 2013, 103 p.
20. **Ivar Zekker.** Enrichment of anaerobic ammonium oxidizing bacteria for nitrogen removal from digester effluent and anammox process acceleration by intermediate compounds. Tartu, 2013, 142 p.
21. **Annika Uibopuu**. Communities of arbuscular mycorrhizal fungi in spruce forest ecosystem and their effect on performance of forest understorey plant species. Tartu, 2013, 104 p.
22. **Jekaterina Jefimova.** Leaching of polycyclic aromatic hydrocarbons (PAHs) and heavy metals from the oil shale processing wastes and from waste-based products. Tartu, 2015, 184 p.
23. **Teele Ligi**. Bacterial community structure and its genetic potential for nitrogen removal in the soils and sediments of a created riverine wetland complex. Tartu, 2015, 127 p.
24. **Kuno Kasak**. Greenhouse gas emissions and water treatment efficiency in subsurface flow filters using various substrates. Tartu, 2016, 128 p.
25. **Martin Ligi**. Application of close range remote sensing for monitoring aquatic environment. Tartu, 2017, 146 p.
26. **Mikk Espenberg.** Impact of management on peatland microbiome and greenhouse gas emissions. Tartu, 2017, 152 p.