DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
340

MÄRT ROOSAARE

K-mer based methods for the identification of bacteria and plasmids

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

340

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

340

MÄRT ROOSAARE

K-mer based methods for the identification of bacteria and plasmids

1632 UNIVERSITY OF TARTU
Press

Institute of Molecular and Cell Biology, University of Tartu, Estonia

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Gene technology on May 8, 2018 by the Council of the Institute of Molecular Cell Biology, University of Tartu.

Supervisor: Prof. Maido Remm, PhD
Chair of Bioinformatics, Institute of Molecular and Cell Biology,
University of Tartu, Tartu, Estonia

Reviewer: Prof. Ain Heinaru, PhD
Chair of Genetics, Institute of Molecular and Cell Biology,
University of Tartu, Tartu, Estonia

Opponent: Prof. Ole Lund, PhD
Department of Bio and Health Informatics, Technical University
of Denmark, Kgs. Lyngby, Denmark

Commencement: Room No. 105, 23B Riia St., Tartu, on August 29, 2018, at 10:15.

The publication of this dissertation is granted by the Institute of Molecular and Cell Biology at the University of Tartu.

This research was funded by the European Union through the European Regional Development Fund through Estonian Centre of Excellence in Genomics and Translational Medicine (project No. 2014-2020.4.01.15-0012) and by the Estonian Ministry of Education and Research (institutional grant IUT34-11).

European Union
European Regional
Development Fund

Investing
in your future

ISSN 1024-6479
ISBN 978-9949-77-763-1 (print)
ISBN 978-9949-77-764-8 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, referred to in the text by Roman numerals (Ref. I to Ref. III):

I Kõiv V, **Roosaare M**, Vedler E, Kivistik PA, Toppi K, Schryer DW, Remm M, Tenson T, Mäe A. (2015). Microbial population dynamics in response to Pectobacterium atrosepticum infection in potato tubers. Scientific Reports, 5 (11606), 1–18.10.1038/srep11606.

II **Roosaare M**, Vaher M, Kaplinski L, Möls M, Andreson R, Lepamets M, Kõressaar T, Naaber P, Kõljalg S, Remm M. (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ 5:e3353; DOI 10.7717/peerj.3353.

III **Roosaare M**, Puustusmaa M, Möls M, Vaher M, Remm M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ 6:e4588; DOI 10.7717/peerj.4588.

The publications listed above have been reprinted with the permission of the copyright owners.

My contributions to the listed publications were as follows:

Ref. I Performed the analysis of bacterial 16S rRNA data, prepared the bacterial heatmap and rarefaction figures and participated in the writing of the manuscript.

Ref. II Participated in the development of the StrainSeeker software, created the web tool, performed the analysis and wrote the manuscript.

Ref. III Developed the PlasmidSeeker software, performed the analysis and wrote the manuscript.

LIST OF ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
bp	Base pair
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DDBJ	DNA Data Bank of Japan
EBI	European Bioinformatics Institute
kbp	1,000 base pairs
MALDI-TOF	Matrix assisted laser desorption ionization time-of-flight
MGS	Metagenomic shotgun sequencing
MLST	Multilocus sequence typing
MS	Mass spectrometry
NCBI	National Center of Biotechnology Information
OTU	Operational taxonomic unit
PFGE	Pulsed field gel electrophoresis
PCR	Polymerase chain reaction
RDP	Ribosomal Database Project
SNP	Single nucleotide polymorphism
WGS	Whole genome sequencing

INTRODUCTION

Microbes are found virtually everywhere on Earth, from the ocean floor and deep mines to hot springs, even on our skin and in our gut. They are essential to life on our planet as the primary source for nutrients and also the primary recyclers of dead matter. Many industries harness bacteria, such as food production, biotechnology, medicine and agriculture.

Still, some of the bacterial species can be pathogenic and cause diseases, from mild diarrhea to life-threatening conditions like sepsis. For example, the Black Death pandemics in the Middle Ages were caused by the bacterium *Yersinia pestis.* Nowadays, antibiotics help us against pathogenic bacteria, but a new threat is looming – widespread antibiotic resistance. This is partly facilitated by plasmids, extra-chromosomal DNA sequences readily transferable between bacteria. Plasmids often encode antibiotic resistance genes, which makes them beneficial to bacteria.

This large phenotypic variety among bacterial species raises interest in their identification, either for their commercial potential or pathogenicity. The growing number of sequenced microbial genomes in public databases has provided an invaluable resource for comparative genomics, microbiome research, genetic engineering (Clustered Regularly Interspaced Short Palindromic Repeats, CRISPR/Cas system) and clinical microbiology, to name a few. However, the progress is hindered by the huge amount of complex data. Most of the raw data produced by various studies is in the form of millions of short reads, 50–300 base pairs long. This is because the widely used second-generation sequencing technology (also referred in literature as "next-generation" or "next-gen") has short read lengths and novel single-molecule sequencing approaches are still in active development. Precisely identifying bacterial strains from short sequencing reads is a difficult task, albeit a necessary one for many applications.

The first part of the thesis provides an overview of methods used to identify bacteria. Due to the different nature of the methods used, identification of isolated bacteria and identification of bacteria from environmental samples is presented separately. The second major topic focuses on bacterial plasmids – their importance and a review of methods of the identification of plasmids, both sequencing-based and other methods.

In the research part of the thesis I describe *k*-mer based methods for the identification of microbes: (i) analyzing 16S rRNA gene sequencing data to reveal microbial community dynamics in potato tubers in response to infection with *Pectobacterium atrosepticum*; (ii) developing a novel algorithm for strain-level identification of bacteria from whole genome sequencing data without assembly and (iii) developing a tool for the detection of known plasmids from bacterial whole genome sequencing data without assembly.

1. REVIEW OF THE LITERATURE

1.1. Identification of bacteria – a history

Identification of bacteria is one of the cornerstones of microbiology and critical in many areas, from food safety monitoring to clinical diagnosis (Emerson et al., 2008; Janda and Abbott, 2002). Identification *per se* comprises both the discovery of novel organisms and the detection of known bacteria. Thanks to a multitude of advances in sequencing, microbiology and bioinformatics, there is now a wealth of information about most of the common bacterial species, available to everyone through public databases maintained by the National Center of Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ). In case of common bacteria, the identification process nowadays often consists of finding just the closest match from a database.

Historically, coming up with suitable taxonomic classification schemes has been challenging in case of bacteria as the classical definition of species as "a group of organisms that can interbreed and produce fertile offspring" does not apply to them (Emerson et al., 2008). Generally, bacteria can be classified according to their phenotypic traits, genome sequence or a combination of both. Early studies dating back to the 19th century separated bacteria into groups based on their morphology, size and motility (Janda and Abbott, 2002). Later, biochemical reactions were used, such as gram staining. Combining different biochemical tests into a single kit gave rapid and accurate results that could also be used in clinical microbiology. As an example, the API 20E test strip, invented in the 1970-s, consists of 20 various biochemical tests and is still in use (Janda and Abbott, 2002). With the advent of DNA sequencing, genotyping methods were rapidly developed, ranging from bacterial 16S rRNA gene analysis to whole genome sequencing (WGS). Still, there is much controversy in what exactly constitutes a bacterial species. This problem is exacerbated by the fact that bacterial species have a set of genes common in most of the strains (core genome) and individual strains also have their own unique set of genes (constituents of accessory genome) (Konstantinidis et al., 2006; Rouli et al., 2015).

Bacterial identification can be broadly divided in two, based on the type of the sample. Firstly, bacteria can be isolated and grown in a culture, so that a sample would only consist of clones from the isolated bacterial strain. The second option is taking a sample straight from the environment, so it contains all representative species of a certain habitat. Environmental samples may contain anywhere from a few to hundreds of different bacterial species, viruses and other organisms (Wooley et al., 2010), making them much more complex to analyze compared to isolated bacteria. Identification techniques applicable for bacterial isolates often cannot be used in case of environmental samples, therefore the identification of bacteria from environmental samples is reviewed as a separate topic.

1.2. Identification of isolated bacteria

In case of isolated bacteria, we can assume that the sample consists of clones. Provided no contamination is present, all the DNA, RNA, proteins and biochemical traces should point to a single strain, simplifying the identification process. However, there are some disadvantages of studying isolated bacteria, most being tied to the culturing process itself. Firstly, only a small fraction of bacterial species can be cultured in laboratory settings (Wooley et al., 2010). Secondly, the culturing step takes extra time, which can be critical when dealing with pathogenic bacteria. Culturing times can range from 12 hours for rapidly growing bacteria to more than two weeks for *Mycobacterium tuberculosis* (Bradley et al., 2015).

The methods used to identify isolated bacteria can be divided into two main categories – either phenotypic or genotypic. Phenotypic methods deal with morphology, biochemical makeup and metabolic attributes while genotypic techniques are based on profiling the genetic material (primarily DNA) of the bacterium. Genotypic methods are not affected by the state of the organism, the phase of growth or the culturing medium (Emerson et al., 2008). They can be either based on sequencing or detecting specific DNA profiles or fingerprints.

1.2.1. Phenotypic methods

Phenotypic methods help to distinguish between bacteria based on their physical properties (size, motility), metabolic properties (such as an ability to degrade lactose), biochemical makeup (proteins produced) or any other attribute that could be readily measured or detected (Emerson et al., 2008). Various methods offer different identification resolutions. Gram staining, for example, can only provide a very broad level of classification as it is based on just a single attribute – the composition of bacterial cell wall. On the other hand, proteomics – analyzing all the proteins produced by a bacterium – offers a multitude of different reference points, enabling species-level classification. Most proteomics tools are based on mass spectrometry (MS), a technology that separates molecules based on their mass-to-charge ratio. Among these, matrix-assisted laser desorption/ionization time-of flight (MALDI-TOF) MS is the most commonly used to identify bacterial species, mainly because of its reproducibility and the ability to analyze whole bacterial cells directly (Emerson et al., 2008). There are specialized platforms for clinical microbiology use, such as the MALDI Biotyper Systems by Bruker.

Every additional step in the identification process takes extra time, which can be critical in case of severe conditions caused by pathogenic bacteria, such as sepsis. Compared to a few hours at best for DNA sequencing (Fournier et al., 2014), MS is several times faster, taking only minutes. It has been shown that MALDI-TOF MS can accurately identify bacteria directly from blood cultures of patients with sepsis and the information obtained could have been used to administer appropriate antibiotics earlier, improving the outcome for a patient in many cases (French et al., 2016).

1.2.2. DNA fingerprinting methods

DNA can be analyzed as a sequence of nucleotides, but also by breaking it into multiple shorter fragments and analyzing their length and abundance. Organisms that are closely related to each other usually produce similar or identical fragment profiles or fingerprints. However, it is important that these fragments are not created randomly, but in a certain, reproducible manner, so that a reference database could be used to compare fingerprints and find the closest match. To date, two main approaches have been used.

The first is to design a set of primers and use polymerase chain reaction (PCR) to amplify certain sections of DNA, such as repetitive elements. Multiplex PCR can be used to amplify more than a single region or to test for several different species.

The second option is to use restriction enzymes that cleave DNA at recognition sites specific to the enzyme and separate the fragments by pulsed-field gel electrophoresis (PFGE) (Emerson et al., 2008). For the fingerprinting methods to be useful, large and comprehensive databases are necessary as a profile itself does not give much information. In case of PFGE, the United States of America has a country-wide network for sharing and using PFGE data called PulseNet, which has already been operating for more than 20 years. Its many success stories highlight the importance of standardization and effective data sharing between laboratories (Boxrud et al., 2010).

1.2.3. 16S rRNA gene and multilocus sequence typing

First sequencing experiments involving the 16S rRNA gene took place more than four decades ago and sequence comparisons have shown that 16S rRNA genes are highly conserved on the species/genus level, but differ on higher taxonomic levels (Woo et al., 2008). This made the 16S rRNA gene very useful for both phylogenetic and identification purposes. 16S rRNA analysis brought with it a paradigm shift as prokaryotes were divided into different domains – bacteria and archae (Woese and Fox, 1977).

Since 1990s, 16S rDNA sequencing has been widely used for routine bacterial identification, especially in case of species for which phenotypic methods were unreliable (Janda and Abbott, 2002; Woo et al., 2008). A characteristic that makes the 16S rRNA gene so amenable for analysis is that every bacterial genome has at least one copy of the gene (Klappenbach, 2001). PCR can be used to amplify variable regions in the 16S rRNA gene, which are sequenced afterwards (Land et al., 2015). For sequence comparison, there are several publicly available 16S rRNA gene databases such as GreenGenes (http://greengenes.secondgenome.com) (DeSantis et al., 2006), SILVA (https://www.arb-silva.de/) (Quast et al., 2013) and Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu) (Cole et al., 2014). However, as 16S rDNA is very similar in closely related species, for example the *Mycobacterium* genus, it does not help to differentiate between them, even as it may be necessary for making clinical decisions (Woo et al., 2008). For more precise, strain-level identification, other supplementary methods are necessary.

Multilocus sequence typing (MLST) approach is similar to 16S rDNA sequencing, but several different genes are used instead of just a single one, creating more combinations and thereby increasing the identification resolution. Originally developed as a tool in clinical microbiology, it has been used also in bacterial population genetics. Each MLST scheme consists of several MLST alleles – fragments of highly conserved housekeeping genes about 350–600 bp long (Maiden, 2006). In order to facilitate MLST analysis, the number of loci is kept at a minimum level, usually just enough to distinguish between pathogenic and non-pathogenic strain groups. Most MLST schemes use between 6–10 loci, depending on the bacterial species (Maiden, 2006). A unique combination of MLST alleles is called a sequence type and knowing it can provide valuable information. For example, *Escherichia coli* with sequence type 131 is a highly virulent type, responsible for most urinary tract and bloodstream infections (Petty et al., 2014).

In order to obtain a sequence type for an isolate, there are different approaches. The traditional way was to first use PCR to amplify MLST alleles and then apply Sanger sequencing on the amplicons (Larsen et al., 2012). As sequencing prices have decreased, WGS is nowadays a more feasible option as it can also shed light on the resistance genes and many other attributes that can be obtained from the full genome sequence. Online tools with a graphical user interface, such as the MLST server (www.cbs.dtu.dk/services/MLST) (Larsen et al., 2012), make it possible to identify the sequence types of strains of interest without bioinformatics skills. The largest and most used online resource for MLST schemes is the PubMLST database (http://pubmlst.org) (Maiden, 2006).

1.2.4. Whole genome sequencing

Bacterial whole genome sequencing, as the name suggests, signifies the sequencing of the full-length bacterial genome, instead of a set of specific markers or regions. It is also called "whole genome shotgun sequencing", as most of the sequencing approaches involve randomly fragmenting long chromosomal sequences of isolated bacteria, which are then sequenced and assembled into longer stretches known as contigs (Kwong et al., 2015). WGS represents the ultimate step in the bacterial genetic analysis as it captures everything in the genome with the exception of epigenetic markers. On top of the taxonomic label, WGS can give a wealth of other information about the isolate, such as its antibiotic resistance genes and plasmids.

In 2008, Emerson et al. discussed whether one day it would be feasible to just sequence the genome of an isolate instead of using MLST, the 16S rRNA gene analysis or phenotypic tests as means for the identification (Emerson et al., 2008). A decade later, WGS is now much more widespread and the costs associated with WGS are have come down, thanks to the rapid advancement of high-throughput second-generation sequencing technologies (Land et al., 2015). The number of sequenced prokaryotic genomes in the NCBI GenBank database has almost doubled every year since 2010 (Land et al., 2015), from less than 1,000 genomes to almost 120,000 in 2017. The problem has shifted from the lack of data to an overabundance of data.

Along with the very welcome reduction in price, second-generation sequencing technologies also reduced the length of reads. Sanger sequencing provided reads up to 1000 bp in length, but read lengths produced by the commonly used Illumina MiSeq platform range from 150 to 300 bp (Kwong et al., 2015; Land et al., 2015). Assembly of these short reads is a complex task and high sequencing coverage is needed to get a complete bacterial genome without gaps, which again raises the overall cost. Because of this, it is often feasible just to produce draft genomes as they still contain most of the usable information, even as they consist of many contigs with gaps in between (Mavromatis et al., 2012).

In the last several years, third-generation sequencing technologies have also become more widespread. Third-generation refers to single molecule sequencing, in which the DNA is not fragmented, but analyzed as a continuous strand, enabling unprecedented read lengths of up to 150 kbp (Jain et al., 2016). The flagship of third-generation sequencing solutions is currently Oxfond Nanopore Technologies, but nano-scale devices made of graphene are also showing promise (Heerema and Dekker, 2016). Nanopore devices are based on biological pores that are able to detect the change in electrical current as the DNA passes through the pore. MinION, the most popular third-generation sequencer, is the size of a large USB stick and enables live transmission of sequencing data, which means that reads can be analyzed even as the sequencing process in still underway (Jain et al., 2016). However, MinION has its own drawbacks, the most serious one being inaccurate base calling due to the rapid pace of the DNA passing through pores. Initially, only 66% of called bases were successfully aligned to matching bases in reference sequences (Jain et al., 2015). This has now been improved to 92% (Jain et al., 2017). Still, this is low compared to 99.6–99.8% that the widely used Illumina sequencing platforms can achieve (Schirmer et al., 2016).

1.2.5. Assembly-based vs assembly-free methods for WGS data analysis

First, it is necessary to introduce the *k*-mer, a ubiquitous concept in the field of modern bioinformatics. Simply put, a *k*-mer is an oligomer of DNA or other biological sequence with a length of *k* bases (Figure 1). Any long sequence, ranging from reads to the whole genome of an organism, can be represented as a list of *k*-mers and their respective abundances. *K*-mers have been employed by a multitude of bioinformatics tools: Basic Local Alignment Search Tool (BLAST) uses them for seed alignment search (Altschul et al., 1997), they are represented in de Bruijn graphs that many read assemblers are based on (SPAdes, Velvet) (Bankevich et al., 2012; Zerbino and Birney, 2008), shotgun metagenomics data analysis tools (Kaiju, Kraken) (Menzel et al., 2016; Wood and Salzberg, 2014) and for phylogenetic distance approximation (Ondov et al., 2016).

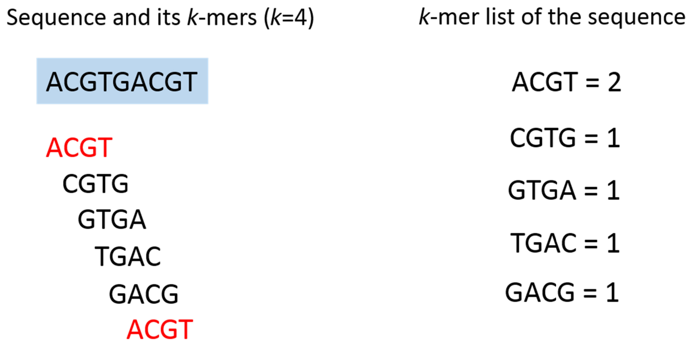

Figure 1. DNA sequence and its *k*-mers. On the left, a DNA sequence is divided to *k*-mers with a sliding window of *k* nucleotides (*k*=4). Each *k*-mer overlaps the previous and the next *k*-mer by *k*-1 nucleotides. Identical *k*-mers are highlighted in red. On the right, the corresponding *k*-mer list is shown, which consists of the sequence and the abundance information of each *k*-mer.

Even with the advent of single-molecule sequencing, most WGS data still consists of 100–300 bp Illumina reads, due to the low cost per bp and widespread use of the technology (Schirmer et al., 2016). However, getting the necessary information from short reads is complicated as they are considerably shorter than MLST loci, antibiotic resistance genes and many other regions of interest in bacterial genomes. Therefore, WGS data analysis approaches can be broadly divided in two – tools that require assembly and tools that can work on raw, unassembled reads. Assemblies are generally required to annotate novel strains (Edwards and Holt, 2013) while the identification of known strains and their genes can be done without an assembly (SRST2, KvarQ) (Inouye et al., 2014; Steiner et al., 2014).

Assembly is the process in which overlapping short reads are put together into contigs. On the one hand, contigs simplify the downstream analysis as they often span several kbp in length (Bankevich et al., 2012) and could be analyzed even with local alignment search tools such as BLAST. On the other hand, creating a high quality assembly is a complex process, often involving pre-processing of reads, optimization of parameters for the assembly program and comparing different results to find the best one (Inouye et al., 2014). Also, assemblers are strongly affected by the choice of parameters and the nature of the data, making it harder to pick the right option for the task at hand (Koren et al., 2014).

There is a variety of different tools available for prokaryotic genome assembly. Some are able to assemble genomes *de novo*, like SPAdes and Velvet. However, if high-quality genome sequences of similar strains are available, they can be used as a reference to facilitate the assembling process, such as the Ragout tool does (Kolmogorov et al., 2014). The quality of an assembly ranges from a draft genome made of many contigs separated by gaps to a complete genome, in which a chromosome is captured in a single contig. Multiple contigs of draft genomes are usually ordered against a close reference genome (Edwards and Holt, 2013).

A genome assembly can be used for many purposes. An important part of genome annotation is the identification of genes. While it is feasible to detect known genes or mutations even from unassembled WGS reads (Inouye et al., 2014; Steiner et al., 2014), *ab initio* gene prediction usually requires assembled reads (Delcher et al., 2007). Also, an assembly can give information about the layout of the genes on the chromosome.

For decades, sequence analysis has mainly been done by comparing an unknown sequence to known ones by arranging them in a way that similar parts of the sequences are next to each other. This is called sequence alignment. For a long time, alignment of sequencing reads to a reference genome or with each other has been a prerequisite of genome-level sequence analysis. Nowadays, the general trend is towards *k*-mer based, alignment-free sequence analysis methods as they are much faster, but provide a level of accuracy similar to their counterparts that require assembling (Bradley et al., 2015; Gupta et al., 2017; Inouye et al., 2014). Compared to the complex sequence alignment process which involves the selection of a suitable scoring matrix and finding the optimal alignment out of many possible alignments, the exact matching of *k*-mers is very fast as it is essentially a yes–no question.

As an example, we can take the tasks of MLST and the prediction of antibiotic resistance, both of which give important information about a bacterial isolate. Due to the low cost of WGS, MLST is now increasingly performed on WGS data as it is cheaper than the traditional MLST described above. The MLST server web tool assembles user-submitted WGS reads and identifies the sequence type using a BLAST search (Larsen et al., 2012). A novel program, stringMLST, achieves the same goal, but does not require assembly and is thereby several times faster, without any loss in accuracy (Gupta et al., 2017). Instead of searching for alignments between the input data and MLST loci, StringMLST counts *k*-mers that are present in any of the MLST alleles and the final sequence type is derived from alleles with highest number of *k*-mer hits.

The detection of antibiotic resistance is important for both surveillance and effective clinical treatment purposes. Along with the traditional antibiotic disk-based methods, *in silico* phenotype prediction from WGS data is becoming more widely adopted. ResFinder is a web tool that identifies antibiotic resistance genes in a manner similar to MLST server – the input WGS reads are assembled and queried with BLAST against a curated database of antibiotic resistance genes (Zankari et al., 2012). A few years later, the authors of ResFinder developed KmerResistance, a tool that examines the co-occurrence of *k*-mers between the WGS data and a database of resistance genes and is both faster and more accurate than ResFinder (Clausen et al., 2016). Another tool, KvarQ, can also detect known single nucleotide polymorphisms (SNPs), including phylogenetic markers and resistance-causing mutations (Steiner et al., 2014). Both KvarQ and KmerResistance are based on *k*-mers and do not require an assembly.

1.2.5.1 Strain identification from WGS data

However, there are few assembly-based or assembly-free methods specifically designed for precisely identifying the bacterial isolate itself. The 16S rRNA gene sequence resolution is generally limited to species level (Woo et al., 2008) and so are tools that use the 16S rRNA gene as the marker (Saputra et al., 2015). MLST is more discriminative, but even strains with the same sequence type may have differences of thousands of SNPs (Petty et al., 2014).

One solution is to design custom probe sequences targeting a panel of genes other than the the 16S rRNA gene (Bradley et al., 2015). While this can be effective for a few chosen species, it is very time-consuming to construct and validate the probes for all clinically relevant bacteria. Another option is to determine the phylogenetic relationship of the isolate in regard to other strains of the same species, as was done by Petty et al. with *E. coli* sequence type 131 strains (Petty et al., 2014). This can distinguish between very similar strains, but it needs high sequencing coverage and read assembling. Tools meant for identifying bacteria from metagenomic shotgun sequencing (MGS) samples (mentioned below in detail) can be also used for identifying isolates as most of them assign a taxonomic label to each read (Saputra et al., 2015).

1.3. Identification of bacteria from environmental samples

Studying bacteria in different habitats can give us valuable information and answer many important questions – which bacteria are there, what are they doing, how do they interact with each other and their environment. Bacteria are present almost anywhere, from the human gut to the ocean floor. In order to get an unbiased view on bacterial communities, they have to be studied directly from environmental samples because many bacterial species cannot be cultivated in the laboratory conditions and most bacterial communities do not consist of a single species (Wooley et al., 2010). Due to the fact that environmental samples require no culturing, identifying bacteria directly from environmental samples also holds great promise for the field of clinical microbiology, especially in the cases where the infection is severe and requires immediate action (Hasman et al., 2014). Compared to bacterial isolate samples, which contain clones of a single strain, environmental samples can be much more complex and contain bacteria from many different taxa (Huttenhower et al., 2012).

Phenotypic methods that are well suited to identify bacterial isolates by their specific DNA or peptide fingerprints, such as PFGE and MALDI-TOF MS, cannot be used to identify bacteria from environmental samples as the fingerprints can originate from multiple bacteria, muddling the overall sample profile. Therefore, most of the studies have analyzed DNA sequences to shed light on bacteria in complex samples, hence the term "metagenomics" as the study of sequence data obtained directly from the environment (Wooley et al., 2010).

There are two main approaches for environmental sample sequencing. The first option is based on the assumption that every bacterium has at least one copy of the 16S rRNA gene and the respective sequences can be amplified with PCR using universal primers. Then, the 16S rRNA gene amplicons are sequenced and analyzed (Wooley et al., 2010). Due to the length of the 16S rRNA gene (~1.5 kbp), only a few variable regions in the gene are usually sequenced (Li et al., 2012). The second option is to use shotgun sequencing, in which the whole DNA is extracted from all organisms, sheared into fragments and sequenced. Both approaches have their strengths and weaknesses and numerous tools have been published to facilitate the analysis of either type of data.

1.3.1. 16S rRNA gene sequencing

The analysis of the 16S rRNA gene sequence data usually begins with the preprocessing of reads – filtering low-quality bases and removing adapter sequences used in the PCR amplification step. Afterwards, sequences are clustered into operational taxonomic units (OTU), conventionally with a 97% similarity threshold, with each OTU having its own consensus sequence (Li et al., 2012). This is done in order to remove errors caused by sequencing, which would otherwise create many false-positive 16S rRNA gene sequences. Many clustering tools have been published, CD-HIT-OTU being one of the better supported ones, which also has a web server available (Li et al., 2012). The 16S rRNA gene databases mentioned above can be used to identify OTU consensus sequences and bioinformatics pipelines like QIIME (Caporaso et al., 2010) to visualize the final results. Also, there are packages consisting of several programs to cover all the steps of the 16S rRNA gene analysis (Schloss et al., 2009).

Compared to MGS, the 16S rRNA gene sequencing is more sensitive in detecting species with low abundance, both because of the PCR amplification step and the fact that not all MGS reads are specific to a bacterial species. Overall, the 16S rRNA gene sequencing provides a robust approach for species-level bacterial identification from environmental samples and has been shown to correlate well with MGS (Zhernakova et al., 2016). However, when calculating abundance values for OTUs it must be kept in mind that the copy number of the 16S rRNA gene in different species can range from 1 to 15 (Klappenbach, 2001), which could lead to biased results if not taken into account.

1.3.2. Metagenomic shotgun sequencing

Compared to the 16S rRNA gene sequencing, MGS can potentially give much more information as all the DNA in the sample is randomly sequenced, not only a single gene. However, this also means that the sensitivity is worse as there may be only a few reads from species with lower abundances, which in turn may not contain enough information for accurate classification. MGS data can be very complex as reads are often short and from random genome locations of an unknown number of organisms. As with WGS, the first question is whether to assemble the reads or not. For simple taxonomic profiling or even functional analysis, it is often unnecessary, but required for deeper insights into the community, such as connecting specific metabolic functions to certain bacterial taxa. This is because a single short read does not usually contain both the gene of interest and taxonomical info. However, assembling MGS data is more complex than WGS data as reads originate from many organisms.

Like the 16S rRNA gene sequencing, MGS is used for taxonomical profiling of samples, answering the “who is there?” question. However, functional analysis of the community or the “what are they doing?” question can only be found out by MGS. By examining the repertoire of genes represented in a MGS sample, we can predict functions and pathways represented in the community, such as photosynthesis and metabolism of various compounds (Silva et al., 2015).

1.3.3. Bioinformatics tools for taxonomical profiling

Most of the bioinformatics tools developed to identify bacteria from unassembled MGS data assign a taxonomic label to each read separately, using a reference database. The assignment can be done in different ways. First, a read can be aligned to all reference sequences and assigned the label of the best match. BLAST is a well-known and still very widely used method to identify a sequence by finding the best alignment from a database of sequences (Altschul et al., 1997). Second approach is mapping, which is faster than BLAST alignment (Truong et al., 2015) due to novel algorithms being used (Li and Durbin, 2009). Third option, now increasingly used, is *k*-mer based tools.

Interpolated Markov Models (Brady and Salzberg, 2009) and Bayesian statistics (Rosen et al., 2008) have also been employed for the classification of MGS reads and although they have shown very good sensitivity and precision, their classification speed is many times slower than BLAST, making them unusable for large data sets. In the recent years, several studies have benchmarked many different MGS analysis tools, aiding potential users in selecting the right tool for the job (Lindgreen et al., 2016; Peabody et al., 2015).

1.3.3.1 Alignment and mapping-based tools

BLAST has been shown to be a sensitive and precise method for identifying bacteria from MGS data. Numerous MGS analysis tools have relied on BLAST for read identification. A good example is MEGAN, developed in 2007, that required reads aligned by BLAST (or a similar tool) (Huson et al., 2007). However, the number of bacterial species in public databases has been growing constantly and with sequencing instruments generating larger amounts of data, BLAST has become too slow for practical use. It is able to identify only 5–10 thousand reads per minute, whereas MGS samples contain millions of reads and would take many hours or days to analyze. This had led to various other algorithms and heuristics to speed up the search, namely reducing the database size and using more efficient aligners.

A well-known MGS data analysis tool, MetaPhlAn, was used by the Human Microbiome Project consortium (Huttenhower et al., 2012). MetaPhlAn still makes use of BLAST, but in order to speed up the identification process, a reduced reference database is used, thereby decreasing the search space (Segata et al., 2012). Instead of full-length bacterial genomes, only unique clade-specific marker genes are kept in the database. From 2,887 bacterial genomes available at the time, the authors selected 400,411 genes that best represented each taxonomic unit. The search process consists of mapping reads to the marker genes in the database and taxon abundances are based on the read coverage of each marker. Due to the smaller database size, MetaPhlAn is 50–100x faster than BLAST (Segata et al., 2012; Wood and Salzberg, 2014).

The authors of MetaPhlAn have since augmented their program with more than half a million markers to support the identification of many more bacterial species and also viruses and eukaryotic microbes. The new tool, MetaPhlAn2, is also able to identify bacteria on the strain level and uses a faster mapping tool for better performance (Truong et al., 2015). Also, the authors of MEGAN have published a novel version of the tool that uses a novel aligner, which is 20,000x faster than translated nucleotide to protein BLAST (BLASTX) and also able to perform functional analysis of samples (Huson et al., 2016).

Parallel computing is another option to speed up the analysis. Read alignment/mapping is usually the most time-intensive part of the analysis. A powerful computing cluster with many CPU cores may reduce the time spent more than a hundredfold (Ahn et al., 2015). Most of the *k*-mer based tools described below can be also parallelized, usually in the *k*-mer counting step.

1.3.3.2 K-mer based tools

A widely adopted approach is searching for exact matches of short k-mers instead of searching for regions with 1–2 mismatches or searching for the best scoring alignment. This greatly improves the search speed as alignment is a complex process that involves searching for an initial seed, extending it and calculating the scores (Altschul et al., 1997). In contrast, exact matching is just detecting the presence or absence of a search string. As long marker genes might easily be missed by exact matching due to sequencing errors and random mutations, exact matching is only suitable for short sequences. In most cases, *k*-mers up to 32 bp are used for this purpose as they can take advantage of the 64-bit architecture of computers, requiring less memory.

Kraken is a well-known example of a *k*-mer based metagenome identification tool (Wood and Salzberg, 2014). It breaks every read into its constituent *k*-mers and for every *k*-mer, it finds the last common ancestor of all the genomes that contain this *k*-mer. This is done using the NCBI taxonomy tree. The read is classified as belonging to the taxon supported by the highest number of *k*-mers. The authors used k=31 as the default *k*-mer length. Compared to BLAST, Kraken is more than 100 times faster, even though its sensitivity and precision are similar. CLARK works in a way similar to Kraken, but instead of using a taxonomy tree, it simply classifies a read according to the bacterium with the highest number of matching *k*-mers (Ounit et al., 2015). Compared to MetaPhlAn described above, both Kraken and CLARK are able to identify more reads as MetaPhlAn only looks for specific markers that may not be present in every read (Lindgreen et al., 2016).

The choice of *k*-mer length is always an important question regarding *k*-mer based tools. Shorter *k*-mers are more sensitive and allow more reads to be classified as they are less prone to containing mutations and sequencing errors. On the negative side, they also cause more misclassifications. Longer *k*-mers are more specific, but less sensitive (Kim et al., 2016; Ounit et al., 2015). In most cases, the authors of a tool have tested it on different datasets using various *k*-mer lengths and suggest an optimal “default length” for *k*, which is essentially a tradeoff between specificity and sensitivity.

For rapid access, the reference database is usually stored in RAM, in the form of a *k*-mer list (Wood and Salzberg, 2014). As the number of reference sequences increases, so does the database and its RAM requirements. This can be a constraint, especially for users with access only to a desktop computer, most equipped with 4–16 GB of RAM. The database of Kraken, containing the genomes of ~4,300 prokaryotes, along with human and viral genomes, takes over 93 GB of disk space, compared to 4.2 GB of Centrifuge. Even with the much smaller database, Centrifuge is almost as sensitive and accurate as Kraken, albeit 2x slower (Kim et al., 2016).

In MGS data analysis, a large fraction of reads may remain unclassified. This is because many organisms are not yet sequenced and DNA-level comparisons with reference sequences do not yield any results with significant identity (Menzel et al., 2016). It is known that protein sequences are more conserved than the underlying DNA, due to the redundancy of the genetic code. Also, prokaryotic genomes are densely packed with genes. This has led to the idea of using protein-based markers instead of DNA, as they should be more sensitive (Menzel et al., 2016). Increased sensitivity and ability to classify more reads is especially important in case of complex environmental samples, like the ones taken from the soil, gut or raw sewage, where many reads, often more than 50% of the total, remain unclassified (Menzel et al., 2016; Wood and Salzberg, 2014). The idea of using protein-based markers has been implemented in a tool named Kaiju, which translates MGS reads into all six possible reading frames and searches for maximum exact matches, using its reference protein database. Authors have shown that Kaiju is more sensitive and able to classify considerably more reads than Kraken (Menzel et al., 2016).

1.3.4. Bioinformatics tools for functional profiling

Two main characteristics of an environmental sample are usually its taxonomic composition ("who is there?") and metabolic functions and networks ("what do they do?"). In case of the human microbiome, bacterial communities may differ a lot in their taxonomical makeup, but still have very similar functional profiles, as shown by the Human Microbiome Project (Huttenhower et al., 2012). In contrast, the analysis of coral reef samples has shown that taxonomies remain similar, but functions adapt to local conditions (Silva et al., 2015). As bacterial species may fulfill different metabolic roles depending on the community, metabolic functions cannot be reliably inferred from the species present in the sample and should be discovered separately. Methods for metagenome functional profiling can be broadly broken in two – those that require assembled reads and those that work on raw reads.

The workflow of assembly-free functional profiling tools is similar to the bacterial identification tools described above, but instead of giving a taxonomic label to a read, it is assigned to a metabolic system or process. In order to decrease the search space, most tools use the taxonomic composition of the sample to only look for functions that species detected in the sample can fulfill, based on their repertoire of genes. To find whether a sequencing read contains a gene fragment, BLASTX has been used, for example by the HUMAnN (HMP Unified Metabolic Analysis Network) tool to metabolically profile the Human Microbiome Project data (Abubucker et al., 2012). With the advent of faster protein alignment tools, novel functional profiling tools have been published that facilitate the analysis of large data sets (Silva et al., 2015). For example, RAPSearch2 is about 100 times faster than BLASTX, thanks to a reduced amino acid alphabet of 10 symbols, each representing a group of amino acids (Zhao et al., 2012).

While it is important to identify metabolic functions represented in the bacterial community, the question of which organism is responsible for a specific role remains. Also, as many species in environmental samples may be novel, there may be a number of genes and functions not represented in the reference databases. Short reads contain just gene fragments, making it hard to predict novel genes and connect genes to certain taxa. MGS data assembly requires special assembling tools, which do not assume that all reads are from a single genome. MetaSPAdes (Nurk et al., 2017) is a recently published metagenome data assembler that builds on the SPAdes assembler, but is also based on de Bruijn graphs. Still, even state-of-the-art metagenome assemblers often fail to produce full-length genomes (Kang et al., 2015). As a close substitute to a full genome, contigs are binned – all contigs predicted to belong to a strain (or a group of similar strains) are put together, containing most of the genes of the strain. Binning can be either supervised (based on similarity to reference genomes) or unsupervised (using sequence composition) (Kang et al., 2015). For example, the unsupervised binning tool MetaBAT (Kang et al., 2015) uses tetranucleotide frequencies and the read coverage to cluster contigs with similar composition. The last step is to predict open reading frames and potential genes. For this, many tools are available, such as FragGeneScan (Rho et al., 2010) and Glimmer (Delcher et al., 2007), both of which incorporate Markov models. Also, there is a novel variant of Glimmer (Glimmer-MG) dedicated to handling MGS data (Kelley et al., 2012).

1.4. Bacterial plasmids

Plasmids are extra-chromosomal genetic elements, ranging from 1 to 1000 kbp (Nyberg et al., 2016), which are capable of autonomous replication and transferable between host cells (Orlek et al., 2017). Often, cells contain multiple plasmids in different copy numbers. Plasmids are important vectors of horizontal gene transfer between bacteria and can directly contribute to the dissemination of genes involved in antibiotic resistance and virulence. Such genes may confer phenotypes that are subject to positive selection in the bacterial community, possibly making multidrug-resistant bacteria more prevalent. The rapid emergence of widespread antibiotic resistance (Ventola, 2015) makes the identification of bacterial plasmids an important task. However, this is complicated by the tendency of plasmids to readily gain, lose and rearrange genetic information (Orlek et al., 2017).

1.4.1. Plasmid identification methods

A variety of methods have been used for plasmid detection and identification, all with their own merits and drawbacks. PCR can be used to amplify certain regions in plasmids that are conserved enough to be reliably found from most plasmids. PCR-based replicon typing targets either the conserved replicon sites (Carattoli et al., 2014) or relaxase proteins (Alvarado et al., 2012) of plasmids, and it can be expanded to target many replicons by using multiplex PCR. Also, PCR can be used to detect plasmids from environmental samples (Smalla et al., 2015). Compared to sequencing, PCR is less labor-intensive and faster. However, there are also numerous drawbacks: multiplex PCR is difficult to extend to cover all novel plasmid groups (Carattoli et al., 2014), assays take time to optimize and the result contains no sequence information about the plasmid itself. Also, PCR-based typing requires previous knowledge of the targeted sequence (Müller et al., 2016).

A novel approach to identify intact plasmids is optical DNA mapping. It is based on the visualization of plasmid DNA stretched on a surface or in nanofluidic channels. With the help of fluorescent dyes that bind to either AT or GC-rich regions of DNA, a unique barcode roughly depicting the DNA sequence of a plasmid can be made visible by using fluorescence microscopy. Barcodes can be also calculated *in silico* for known plasmids, simplifying reference database creation (Nyberg et al., 2016). Moreover, optical mapping can be combined with CRISPR/Cas9 to identify various genes located on the plasmid. A guide-RNA, complementary to the gene of interest, cuts the plasmid at the location of the gene, which can be visualized with optical maps (Müller et al., 2016). While overall a very promising approach, optical mapping may not yet be suitable for the detection of short (<50 kbp) plasmids (Nyberg et al., 2016).

As with bacteria, lower sequencing costs and the increasing number of plasmids in the public databases has made sequencing a viable option also for the detection and identification of plasmids (Smalla et al., 2015). Most of the studies have dealt with bacterial WGS data (Orlek et al., 2017; Smalla et al., 2015), MGS data being too complex due to short reads coming from a variety of organisms, some of which may be unknown. WGS data is either assembled or mapped to reference sequences. Read mapping with tools such like SRST2 can help to rapidly detect loci of interest, such as antibiotic resistance genes, but whether they are located on the plasmid or the bacterial chromosome remains unknown (Orlek et al., 2017).

The assembly of plasmids from bacterial WGS reads requires a different approach than assembling bacterial genomes as reads may originate from multiple plasmids and the bacterial chromosome. PlasmidSPAdes (Antipov et al., 2016), based on the SPAdes assembler, uses the read coverage of contigs to distinguish between plasmid and bacterial sequences. The result is given as a list of detected plasmids, each with their respective contigs. As reads are assembled *de novo*, without any reference database, PlasmidSPAdes is able to detect novel plasmids. Carattoli et al. developed the plasmidfinder web tool (Carattoli et al., 2014), which searches for conserved replicon sites using BLAST and compares them to a curated database of plasmid replicons. As a prerequisite, reads must be assembled as the targeted replicon sites are often longer than 300–400 bp. Plasmidfinder is able to detect only plasmids which contain targeted replicons.

2. AIMS OF THE STUDY

The aim of this thesis was to explore the feasibility of *k*-mer based algorithms to identify microbial strains using DNA sequences. We started out with the 16S rRNA gene sequencing data and then focused on bacterial whole genome sequencing data. We set out to solve two important questions – how to provide quick strain-level identification of bacteria based on sequencing data and how to detect and identify any accompanying plasmids. From the start, we concentrated on methods that work on raw, unassembled reads to avoid potential biases related to genome assembly.

The specific aims of this study were:

- To develop a *k*-mer based method capable of identifying bacteria on the strain level from unassembled bacterial whole genome sequencing reads.
- To develop a *k*-mer based method to detect known plasmids from unassembled bacterial whole genome sequencing reads.

3. RESULTS AND DISCUSSION

3.1. Microbial population dynamics in potato tubers (Ref. I)

Although potato is an important food item consumed globally, the studies on the microbial population dynamics within stored potato tubers have been sparse, last one being conducted in 1979. In this study, we shed light on the microbial population dynamics in response to *Pectobacterium atrosepticum* infection in potato tubers.

Experiments were conducted with two batches of potatoes – harvested in 2012 (Experiment 1) and 2013 (Experiment 2). Each potato was infected with *P. atrosepticum* and samples taken from the macerated tissue after two, five and eight days post-infection (Figure 1, ref. I). Bacteria were cultivated from the samples and subjected to the 16S rRNA gene Sanger sequencing. As many bacterial species are not cultivable, several samples from each time-point were chosen for 16S rDNA amplification and sequencing to follow bacterial community dynamics. Sequenced amplicons from Experiment 2 were ~300 bp, spanning the variable regions V1 and V2 of the 16S rRNA gene. Amplicons from Experiment 1 were from the same region, but only ~100 bp long and from random locations. From the cultivated bacteria, a ~1,500 bp long 16S rRNA gene fragment was amplified and Ribosomal Database Project classifier (Cole et al., 2014) was used for species identification.

The identification of bacterial community composition from 16S rDNA amplicons was more complex as they had to be first clustered into OTUs. As reads from Experiment 1 were from random locations, we only clustered sequences from Experiment 2. First, we tried to use the mothur package (Schloss et al., 2009). Its authors were able to analyze a set of 222,000 sequences in a few hours. However, our dataset consisted of more than 4 million sequences and mother was unable to handle it. Therefore, we chose another tool, AbundantOTU (Ye, 2010). It is based on a consensus alignment algorithm, which first searches for an abundant *k*-mer seed and then extends it to form an OTU consensus sequence. Nucleotides for the extension are chosen based on which one would result in the most abundant sequence. After removing chimeric sequences, we got a total of 294 OTUs. Experiment 1 reads were clustered by using BLAST with all 294 OTU consensus sequences of Experiment 2 as the database. Rarefaction analysis with Experiment 2 data showed that deeper sequencing would have given more OTUs, especially in the case of uninfected potatoes (Figure 5, ref. I).

Overall, the results from both experiments showed that *P. atrosepticum* was dominant in the beginning of the infection, but was taken over by resident endophytic bacteria as the infection progresses (Figure 4, ref. I). A reason for this may be that as *P. atrosepticum* breaks down the plant cell wall, it generates a large amount of free sugars. These are consumed by *Enterobacteriaceae* and *Pseudomonadaceae*, which have an advantage in the early phase of the infection. Later, bacteria that specialize in consuming less energetic substances, such as *Comamonadaceae*, took over (Figure 6, ref. I).

3.2. StrainSeeker (Ref. II)

Pathogenic bacteria represent a world-wide problem to human health. Due to lowering sequencing costs, WGS is being increasingly used to identify bacteria. Bioinformatics tools help to analyze large amounts of data generated by WGS and are able to detect clinically relevant alleles and mutations, provided the sample coverage is high enough (Bradley et al., 2015; Inouye et al., 2014). However, for sub-species classification of pathogens from WGS data, the choice of tools is limited as mostly MLST or specially designed probe sequences are used (Bradley et al., 2015). We set out to develop a program that is able to classify bacterial isolates into clonal groups or clades directly from unassembled WGS reads by using clade-specific *k*-mers.

3.2.1. Database, guide trees and search algorithm

To be able to place unknown isolates into clades with known bacteria, we first needed to create a reference database of bacterial strains and determine relationships between them using a guide tree (Figure 1, ref. II). The database is built by recursively moving *k*-mers into parent nodes if they are present in lower nodes. Therefore, *k*-mers of a single strain are spread along the path from root node to the strain. Final database consists of *k*-mer lists specific to each node and strain.

Previously published programs have used the NCBI taxonomy tree (Wood and Salzberg, 2014), but we wanted StrainSeeker to be independent of existing taxonomic systems. Therefore, we decided to allow any Newick-format tree as the guide tree. We downloaded all 4,324 bacterial genome sequences from the NCBI Refseq database, which were available at the time. We could not use the traditional approach of creating the guide tree from a multiple sequence alignment in case of all 4,324 strains due to the lack of common genes able to discriminate between close strains. Instead, we used the *k*-mer based, alignment-free tool Mash (Ondov et al., 2016) to create a matrix of pairwise distances between strains and MEGA6 (Tamura et al., 2013) to build the guide tree of 4,324 strains. We also built two small guide trees, consisting of all 74 *E. coli* strains available from the NCBI Refseq database. The purpose of these was to analyze the effect of different guide trees on the results as one of the small trees was based on a Mash-derived distance matrix and the other was based on a multiple alignment of 126 common *E. coli* genes.

Next step was to create the algorithm that identifies the clade where the unknown isolate belongs to. We used the assumption that the fractions of *k*-mers shared between a strain and each node on the path from the root to the strain on the guide tree are similar to each other and higher than the *k*-mer fractions shared with other nodes. The search algorithm starts from the root of the guide tree and calculates both the observed and expected *k*-mer fractions for each node, the latter calculated from the child nodes (Figure 2, ref. II). The ratio of observed/expected fractions is used to detect where the isolate branches off the guide tree (Figure 3, ref. II).

3.2.2. Performance testing and benchmarking

To find an optimal *k*-mer length and the best guide tree for StrainSeeker, we used a test set of 100 *E. coli* isolates. Their clades were first determined by the "gold standard" approach, building a multiple alignment-based phylogenetic tree that contained both them and 74 *E. coli* strains from Refseq (Figure 4A, ref. II). All strains separated by less than 0.001 nucleotide substitutions per site were considered a clade. Then, we used StrainSeeker to identify the clades of the test strains. Tests were conducted with the large and two smaller guide trees using *k*-mers with lengths varying from 14 to 32 bp. The results showed that the Mash distance matrix-based guide tree of 4,324 strains and *k*=16 were optimal for StrainSeeker, giving the clade-level accuracy 92% (Figure 5C, ref. II). Also, we determined that 25,000 Illumina reads (length 101 bp) are sufficient for clade identification, because the accuracy is not improving with higher sample coverages.

Finally, we benchmarked StrainSeeker against other tools using five samples of different bacterial species. Kraken, Sigma and Reads2Type were chosen for comparisons. All tools except Sigma were based on exact *k*-mer matching and had run times in the range of a few minutes. Sigma is based on read mapping and it spent several hours per sample, illustrating that read alignment is significantly more computationally expensive than exact matching of short *k*-mers (Table 1, ref. II). Clade identification accuracy was only tested in case of Kraken and StrainSeeker, as Sigma was excessively slow and Reads2Type was limited to species level. Compared to Kraken, StrainSeeker was more accurate in determining the clades of 100 *E. coli* isolates. This might be because StrainSeeker does not identify each read separately, but analyzes all the *k*-mers in the sample together. If the exact isolate is not in the reference database, which is usually the case, individual reads may be assigned to multiple different genomes.

To date, StrainSeeker has been used in multiple research projects for identification of strains from either bacterial WGS data or from low-complexity MGS data. Also, StrainSeeker is frequently used for detection of contamination (sequences from another strain or species) in sequencing samples.

3.3. PlasmidSeeker (Ref. III)

Plasmids are double-stranded DNA molecules capable of autonomous replication and conjugation. Bacterial plasmids often carry genes that confer beneficial traits to their hosts, such as antimicrobial resistance or increased virulence. This has directly contributed to the rapid dissemination of multidrug-resistant bacteria. Bacterial WGS data, widely used to identify and characterize bacterial pathogens, also contains sequences from plasmids. However, plasmid detection from bacterial WGS data is complicated as the reads are often short and plasmid sequences may be similar to bacterial sequences. Because of this, plasmid detection tools, such as plasmidfinder and plasmidSPAdes, assemble reads as longer contigs are easier to identify. Our goal was to develop a tool for the detection of plasmids from unassembled bacterial WGS reads, similar to StrainSeeker.

3.3.1. Database, search algorithm and optimal k-mer length

First, we collected all 9,351 available plasmid sequences from the NCBI RefSeq database for our reference database. As some of these sequences were fragments or contained individual genes instead of full plasmids, our final reference set consisted of 8,514 plasmids. The database consists of *k*-mer list files for each reference plasmid and a text-format index file connecting the name of each plasmid to its *k*-mer list. FASTA identifiers are used as plasmid names. The required input for building a database is a multi-FASTA file with plasmid sequences, which is also the format that can be downloaded from the NCBI RefSeq database.

Next, we developed the search algorithm. We considered an approach similar to StrainSeeker, namely finding specific *k*-mers for each plasmid in our reference database. However, this proved unfeasible due to the small size of many plasmids and their high similarity to bacterial sequences. We decided to use all plasmid *k*-mers and compare the median *k*-mer abundances of each tested plasmid to the median *k*-mer abundance of the isolated bacterium. This approach is based on the assumption that the copy number of a plasmid, and therefore its coverage, is different (usually higher) than that of the bacterial chromosome. The search algorithm of plasmidSPAdes is based on the same assumption, but instead of analyzing read coverage of contigs, which requires assembly, we compare *k*-mer abundances to distinguish between plasmid and chromosomal *k*-mers.

A brief overview of the PlasmidSeeker search algorithm:

1. Input sample file (raw WGS reads) is converted to a *k*-mer list, and all *k*-mers that occur only once are discarded, as these are mostly due to sequencing errors.

2. Algorithm finds the approximate genome coverage of the isolated bacterium. For this, a full genome sequence of a reference bacterium, as closely related to the isolate as possible, must be provided by the user.
3. The fraction of detected unique plasmid *k*-mers is found for all reference plasmids. Only reference plasmids with the fraction above a threshold (default 80%) are analyzed further and reported in the output.
4. The average plasmid copy number per bacterial cell is estimated by dividing the median *k*-mer abundance of the given plasmid with the median *k*-mer abundance of chromosomal *k*-mers.
5. Similar plasmids are clustered together in the results. The output is a tab-delimited text file.

For the last part of developing the algorithm, we had to find optimal values for the *k*-mer length and an optimal threshold of the fraction of detected unique plasmid *k*-mers. The latter was necessary because some of the plasmid *k*-mers may be shared with the bacterial isolate and detecting a plasmid *k*-mer might not mean that the plasmid itself is really present in the sample.

As sequences originating from plasmids are distinguished from chromosomal sequences based on their *k*-mer abundances, it is preferable that most chromosomal *k*-mers are unique and not present in any plasmids. Therefore, we analyzed the effect of *k*-mer length on the uniqueness of chromosomal *k*-mers and on the fraction of *k*-mers shared between plasmids and chromosomes (Figure 1, Ref. III). The test showed that *k*-mer length should be at least 20 as shorter *k*-mers have much higher chances of being present in both plasmids and the chromosomal sequence.

Plasmids found in real samples are seldom 100% identical to reference sequences. We assessed how mutations in a plasmid sequence affect the fraction of plasmid *k*-mers detected, using various *k*-mer lengths (Figure 2, Ref. III). Results indicated that longer *k*-mers are less sensitive. Taking all this into consideration, we decided to use *k*=20 as the default value.

To find an optimal threshold of the fraction of detected unique plasmid *k*-mers, we analyzed six bacterial WGS samples, both simulated and real (Table 1, Ref. III). Values of 0.8 and over resulted in no false positives (Figure 4, Ref. III). As higher values decrease sensitivity, we used 0.8 as the default value, meaning that at least 80% of all plasmid *k*-mers must be detected to report it.

3.3.2. Performance testing and benchmarking

To evaluate the performance of PlasmidSeeker, we compared it to plasmidSPAdes. First, we analyzed both simulated and real WGS samples in which the plasmid content was known (Table 1, Ref. III). PlasmidSeeker detected all the correct plasmids and predicted their copy numbers accurately.

Second, we used both tools to detect plasmids from three *E. coli* samples, for which the plasmid content was unknown (Table 2, Ref. III). The tools seem to complement each other as PlasmidSeeker was unable to detect putative plasmids which either had very low copy numbers or were not very similar to reference plasmids. PlasmidSPAdes, on the other hand, failed to detect some of the putative plasmids with high copy numbers.

To sum up, we have developed a novel tool to detect plasmids from bacterial whole genome sequencing data without the need to assemble reads. PlasmidSeeker is suitable to use as a first step in the analysis of plasmid content and it complements tools that assemble reads and are thus able to detect novel plasmids.

CONCLUSIONS

Pathogenic bacteria present a considerable danger to human health. The situation is made worse by the rapid emergence and dissemination of antibiotic resistance, which is partly mediated by bacterial plasmids. Meanwhile, sequencing costs have continuously decreased and WGS is being increasingly used to identify and analyze bacteria.

We developed two *k*-mer based tools for bacterial WGS data analysis, StrainSeeker and PlasmidSeeker. StrainSeeker identifies bacterial strains by assigning them to a clade of the user-provided guide tree. This enables a higher resolution than MLST based identification and is faster than approaches using read mapping. In order to make StrainSeeker accessible also to users without bioinformatics skills, we created a web server with a visual user interface.

PlasmidSeeker detects known plasmids from WGS data by searching for plasmid *k*-mers and comparing their frequency to the frequency of bacterial *k*-mers. As the number of fully sequenced plasmids in public databases is already over 8,000 and growing each year, it is plausible to perform quick monitoring for known plasmids instead of always assembling plasmid sequences *de novo*.

Both tools are able to work with unassembled, raw reads, meaning no preprocessing steps are necessary. Together, they form a comprehensive resource for identifying the isolated bacterial strain and any known plasmids harbored by it, an essential task for both research and clinical purposes.

SUMMARY IN ESTONIAN

K-meeridel põhinevad meetodid bakterite ja plasmiidide tuvastamiseks

Mikroorganismid on meie planeeti asustanud juba miljardeid aastaid ning neid leidub peaaegu kõikjal. Neid on avastatud ookeanisüvikutes olevatest mustadest suitsetajatest, kõrvetavkuumadest allikatest ning sadade meetrite sügavuselt kaevandustest. Isegi meie oleme nendega lahutamatult seotud – baktereid elab nii meie nahal kui ka soolestikus ning nende arv on võrreldav meie enda keharakkude arvuga. Elusloodusе aineringes on mikroorganismidel väga oluline osa orgaanilise aine lagundamises. Paljud tööstusharud kasutavad baktereid oma hüvanguks, rakendused ulatuvad kaevandustes maagi puhastamisest geenide manipuleerimiseni CRISPR/Cas süsteemi abil.

Siiski, bakteritel on ka oma varjukülg – osad neist võivad olla patogeensed ja põhjustada haigusi, kergest kõhulahtisusest eluohtlikeni. Näiteks oli keskajal suure hulga elanikkonnast tapnud Musta Surma põhjustajaks katkubakter *Yersinia pestis*. Tänapäeval aitavad meid bakterite vastu antibiootikumid, kuid järjest suurem probleem on antibiootikumiresistentsuse laialdane levik. Sellele aitavad kaasa plasmiidid – bakterites olevad DNA järjestused, mis on bakteri enda kromosoomist eraldiseisvad ning mida bakterid võivad kiirelt üksteisele edasi anda. Plasmiidid kodeerivad tihti geene, mis annavad resistentsuse mõne antibiootikumi suhtes ning nende omamine võib seetõttu olla bakterile kasulik.

Bakterite tohutu varieeruvus ja nende potentsiaal nii tööstusliku rakendamise osas kui ka haiguste põhjustajatena on tekitanud väga suure huvi bakterite tuvastamise ja määramise osas. Selleks on kasutatud väga palju erinevaid meetodeid, mis jagunevad laias laastus kaheks. Ühed põhinevad bakteri väliste tunnuste analüüsil, nagu näiteks bakteriraku kuju, suurus, selle liikuvus ja erinevad biokeemilised omadused (fenotüüp). Teised meetodid võtavad määramise aluseks bakteri DNA järjestuse (genotüüp).

Viimasel aastakümnel on sekveneerimistehnoloogia väga kiirelt arenenud ning hinnad sedavõrd langenud, et bakteri genotüübi uurimiseks on täiesti mõeldav mitte ainult mõningate DNA-põhiste markerite järjestuse määramine, vaid täisgenoomi sekveneerimine. See on avanud täiesti uued võimalused – näiteks saab ennustada bakteritüve resistentsust erinevatele antibiootikumidele ja kindlaks määrata haiguspuhangute põhjustajaid ning kaardistada nende leviku teid. Uueks probleemiks on aga sekveneerimisandmete analüüs – seninägematult suured andmemahud ning lühikesed lugemid teevad toorandmetest info kätte saamise aeganõudvaks ja keeruliseks. Üheks levinud abinõuks on lugemite assambleerimine ehk kokkupanek pikemateks järjestusteks, kuid see on ajakulukas ning aldis vigadele.

Antud uurimistöö põhiliseks eesmärgiks oli luua bakterite ja plasmiidide tuvastamiseks meetodid, mis ei vajaks eelnevat lugemite assambleerimist ning võimaldaksid töötada sekveneerimiskeskuste poolt toodetud toorandmetega. Ülesande lahendamiseks otsustasime kasutada *k*-meeridel põhinevat analüüsi.

K-meer tähistab lühikest DNA oligomeeri pikkusega *k* nukleotiidi. Pikema DNA järjestuse, näiteks bakterigenoomi, saab jagada lühemateks *k*-meerideks ning vaadelda seda kui *k*-meeride kogumit. Sellise lähenemise eeliseks on sõltumatus lugemi pikkusest – nii pikad kui ka lühikesed lugemid sisaldavad *k*-meere ning analüüsides *k*-meeride hulki, on võimalik määrata algse proovi koostist.

StrainSeeker on meie töögrupis loodud programm bakteritüvede ja liikide määramiseks. Me arendasime välja uudse algoritmi, mis näitab proovis esineva bakteri eeldatavat asukohta kasutaja poolt ette antaval fülogeneetilisel puul. Meie fülogeneetilisel puul põhineva lähenemise üheks suureks eeliseks on see, et uuritav bakter ei pea olema programmi poolt kasutatavas andmebaasis esindatud. StrainSeekeri andmebaas koosneb igale referentsbakterile ja nende gruppidele spetsiifilistest *k*-meeridest. Analüüs põhineb proovis nähtud ning StrainSeekeri andmebaasi põhjal arvutatud eeldatud *k*-meeride hulga suhtel. Me testisime StrainSeekerit saja *Escherichia coli* isolaadi täisgenoomi sekveneerimisandmetega ning tüvede määramise täpsus selles andmestikus oli 92%. Võrreldes teiste programmidega, nagu Kraken ja Reads2Type, oli StrainSeeker täpsem. Lõime ka visuaalse kasutajaliidesega veebiserveri, kus saavad StrainSeekeriga analüüse teostada ka kasutajad, kellel puudub ligipääs arvutusserverile või vajalikud oskused.

Bakterite täisgenoomi sekveneerimisel saadavad andmed sisaldavad tihti ka lugemeid, mis pärinevad bakteris olnud plasmiididest. Plasmiidide tuvastamise ja nende tüübi määramise muudavad keeruliseks nende lühike järjestus ning osaline sarnasus peremeheks oleva bakteri genoomiga. Seetõttu ei õnnestunud StrainSeekeri algoritmi rakendada plasmiidide puhul ning tuli välja töötada uus meetod, mis sai nimeks PlasmidSeeker. Plasmiidset päritolu järjestuste eristamiseks kromosomaalsetest järjestustest kasutasime eeldust, et plasmiidide koopiaarv on tavaliselt suurem bakteri kromosoomi omast, seega võiks ka plasmiidi *k*-meeride keskmine esinemissagedus olla suurem kui bakteri kromosoomi *k*-meeride puhul. Sellise lähenemisega on võimalik bakteritüve täisgenoomi sekveneerimisel saadud järjestustest tuvastada kõiki varasemalt teadaolevaid plasmiide, mida on PlasmidSeekeri andmebaasis kokku 8514. Me testisime PlasmidSeekerit nii simuleeritud kui ka reaalsete bakteri täisgenoomi sekveneerimisandmestikega, millede puhul oli teada proovide tegelik koostis. PlasmidSeeker leidis üles kõik proovides olnud plasmiidid ning määras täpselt ka nende koopiaarvu. Võrdlesime PlasmidSeekerit ka ühe teise programmiga (plasmidSPAdes), mis assambleerib eelnevalt lugemid ja suudab leida ka täiesti uusi plasmiide. Kolme analüüsitud *E. coli* proovi puhul oli teatud osa plasmiide, mille leidsid mõlemad programmid, kuid mõningad plasmiidid leiti vaid ühe programmi poolt. Näiteks ei suutnud PlasmidSeeker tuvastada väga madala koopiaarvuga või andmebaasis olevast referentsist väga erinevaid plasmiide, kuid see-eest tuvastas ta paremini kõrge koopiaarvuga plasmiide.

Kokkuvõttes oleme oma tööga andnud panuse arvutuslikku mikrobioloogiasse, luues uued võimalused bakteriaalsete proovide analüüsiks.

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. *8*.

Ahn, T.H., Chai, J., and Pan, C. (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics *31*, 170–177.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Alvarado, A., Garcillán-Barcia, M.P., and de la Cruz, F. (2012). A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. PLoS One *7*.

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., and Pevzner, P. a. (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics *32*, btw493.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. *19*, 455–477.

Boxrud, D., Monson, T., Stiles, T., and Besser, J. (2010). The role, challenges, and support of pulsenet laboratories in detecting foodborne disease outbreaks. Public Health Rep. *125*, 57–62.

Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., De Cesare, M., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat. Commun. *6*, 018564.

Brady, A., and Salzberg, S. (2009). Classification with Interpolated Markov Models. Nat. Methods *6*, 673–676.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pẽa, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Carattoli, A., Zankari, E., Garcia-Fernandez, A., Larsen, M.V., Lund, O., Villa, L., Aarestrup, F.M., and Hasman, H. (2014). In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob. Agents Chemother. *58*, 3895–3903.

Clausen, P.T.L.C., Zankari, E., Aarestrup, F.M., and Lund, O. (2016). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J. Antimicrob. Chemother. *71*, 2484–2488.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642.

Delcher, A.L., Bratke, K. a, Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiong DNA with Glimmer. Bioinformatics *23*, 673–679.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked

16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Edwards, D.J., and Holt, K.E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inf. Exp *3*, 2.

Emerson, D., Agulto, L., Liu, H., and Liu, L. (2008). Identifying and Characterizing Bacteria in an Era of Genomics and Proteomics. Bioscience *58*, 925–936.

Fournier, P.-E., Dubourg, G., and Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med. *6*, 114.

French, K., Evans, J., Tanner, H., Gossain, S., and Hussain, A. (2016). The clinical impact of rapid, direct MALDI-ToF identification of bacteria from positive blood cultures. PLoS One *11*, 1–9.

Gupta, A., Jordan, I.K., and Rishishwar, L. (2017). stringMLST: A fast k-mer based tool for multilocus sequence typing. Bioinformatics *33*, 119–121.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C.A., Frimodt-Moller, N., and Aarestrup, F.M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J. Clin. Microbiol. *52*, 139–146.

Heerema, S.J., and Dekker, C. (2016). Graphene nanodevices for DNA sequencing. Nat. Nanotechnol. *11*, 127–136.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.J., and Tappu, R. (2016). MEGAN Community Edition – Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. *12*, 1–12.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., Fitzgerald, M.G., Fulton, R.S., et al. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M.B., Pope, B.J., Tomita, T., Zobel, J., and Holt, K.E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. *6*, 90.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat. Methods *12*, 351–356.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 1–11.

Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D. a., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., et al. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Research *6*, 760.

Janda, J.M., and Abbott, S.L. (2002). Bacterial identification for publication: When is enough enough? J. Clin. Microbiol. *40*, 1887–1891.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Kelley, D.R., Liu, B., Delcher, A.L., Pop, M., and Salzberg, S.L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. *40*, 1–12.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Res. *26*, 1721–1729.

Klappenbach, J. a. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic Acids Res. *29*, 181–184.

Kolmogorov, M., Raney, B., Paten, B., and Pham, S. (2014). Ragout – A reference-assisted assembly tool for bacterial genomes. Bioinformatics *30*, 302–309.

Konstantinidis, K.T., Ramette, a., and Tiedje, J.M. (2006). The bacterial species definition in the genomic era. Philos. Trans. R. Soc. B Biol. Sci. *361*, 1929–1940.

Koren, S., Treangen, T.J., Hill, C.M., Pop, M., and Phillippy, A.M. (2014). Automated ensemble assembly and validation of microbial genomes. BMC Bioinformatics *15*, 1–9.

Kwong, J.C., Mccallum, N., Sintchenko, V., and Howden, B.P. (2015). Whole genome sequencing in clinical and public health microbiology. Pathology *47*, 199–210.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics *15*, 141–161.

Larsen, M. V, Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. J. Clin. Microbiol. *50*, 1355–1361.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. Brief. Bioinform. *13*, 656–668.

Lindgreen, S., Adair, K.L., and Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Sci. Rep. *6*, 19233.

Maiden, M.C.J. (2006). Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. *60*, 561–588.

Mavromatis, K., Land, M.L., Brettin, T.S., Quest, D.J., Copeland, A., Clum, A., Goodwin, L., Woyke, T., Lapidus, A., Klenk, H.P., et al. (2012). The Fast Changing Landscape of Sequencing Technologies and Their Impact on Microbial Genome Assemblies and Annotation. PLoS One *7*, 1–6.

Menzel, P., Ng, K.L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. *7*, 1–9.

Müller, V., Rajer, F., Frykholm, K., Nyberg, L.K., Quaderi, S., Fritzsche, J., Kristiansson, E., Ambjörnsson, T., Sandegren, L., and Westerlund, F. (2016). Direct identification of antibiotic resistance genes on single plasmid molecules using CRISPR/Cas9 in combination with optical DNA mapping. Sci. Rep. *6*, 1–11.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. a. (2017). MetaSPAdes: A new versatile metagenomic assembler. Genome Res. *27*, 824–834.

Nyberg, L.K., Quaderi, S., Emilsson, G., Karami, N., Lagerstedt, E., Müller, V., Noble, C., Hammarberg, S., Nilsson, A.N., Sjöberg, F., et al. (2016). Rapid identification of intact bacterial resistance plasmids via optical mapping of single DNA molecules. Sci. Rep. *6*, 1–10.

Ondov, B.D., Treangen, T.J., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

Orlek, A., Stoesser, N., Anjum, M.F., Doumith, M., Ellington, M.J., Peto, T., Crook, D., Woodford, N., Sarah Walker, a., Phan, H., et al. (2017). Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. Front. Microbiol. *8*, 1–10.

Ounit, R., Wanamaker, S., Close, T.J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics *16*, 236.

Peabody, M. a, Van Rossum, T., Lo, R., and Brinkman, F.S.L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics *16*, 363.

Petty, N.K., Ben Zakour, N.L., Stanton-Cook, M., Skippington, E., Totsika, M., Forde, B.M., Phan, M.-D., Gomes Moriel, D., Peters, K.M., Davies, M., et al. (2014). Global dissemination of a multidrug resistant Escherichia coli clone. Proc. Natl. Acad. Sci. U. S. A. *111*, 5694–5699.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. *41*, 590–596.

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

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using N-mer frequency profiles. Adv. Bioinformatics *2008*, 205969.

Rouli, L., Merhej, V., Fournier, P.E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. *7*, 72–85.

Saputra, D., Rasmussen, S., Larsen, M. V, Haddad, N., Sperotto, M.M., Aarestrup, F.M., Lund, O., and Sicheritz-Pontén, T. (2015). Reads2Type: a web application for rapid microbial taxonomy identification. BMC Bioinformatics *16*, 398.

Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N., and Quince, C. (2016). Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics *17*, 1–15.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R. a., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Silva, G.G.Z., Green, K.T., Dutilh, B.E., and Edwards, R. a. (2015). SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. Bioinformatics *32*, 354–361.

Smalla, K., Top, E.M., and Jechalke, S. (2015). Plasmid Detection, Characterization, and Ecology. Microbiol. Spectr. *3*, 1–21.

Steiner, A., Stucki, D., Coscolla, M., Borrell, S., and Gagneux, S. (2014). KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics *15*, 881.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., and Larsen, M.V. (2012). Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. *67*, 2640–2644.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics *28*, 125–126.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science (80-.). *352*, 565–569.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. *30*, 2725–2729.

Truong, D.T., Franzosa, E. a., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

Ventola, C.L. (2015). The antibiotic resistance crisis: part 1: causes and threats. P T A Peer-Reviewed J. Formul. Manag. *40*, 277–283.

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

Woo, P.C.Y., Lau, S.K.P., Teng, J.L.L., Tse, H., and Yuen, K.Y. (2008). Then and now: Use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin. Microbiol. Infect. *14*, 908–934.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Wooley, J.C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. PLoS Comput. Biol. *6*.

Ye, Y. (2010). Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. Proc. – 2010 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2010 153–157.

ACKNOWLEDGMENTS

First and foremost, I wish to thank my supervisor Maido, who welcomed me to the bioinformatics workgroup and ignited my interest in the secrets of programming and bioinformatics. Thank you for your patience, helpful ideas and guidance.

I would not have gotten this far just by myself. I believe that a team is bigger than the sum of its parts and members of our workgroup have helped me tremendously on the journey. Mikk, thanks for being a roommate for all these years, sharing ideas, feedback, ping-pong games and even bread when necessary. Mihkel, supervising you was a great experience and working together with you in many projects that followed even better. Märt – you have clearly shown me the power of statistics. Tarmo, Heleri, Lauris, Kairi, Triinu, Age, Reidar – for the good advice, helping me to improve publications and providing interesting conversation topics, anything from ALU elements to catching pokemons. Also, I have enjoyed the company of colleagues who now work elsewhere – thank you, Tõnu and Mikk! Special thanks go to our lovely secretary Oliivika for handling the never-ending paperwork.

Last but not least – I am very grateful to the nice girls on the other side of the 5th floor – all these board game nights, work lunches and delicious sweets can really help to propel science forward. And of course – Silva, I was lucky enough to become a bioinformatician, meet you and have this adventure of a lifetime together with you!

PUBLICATIONS

CURRICULUM VITAE

Name:	Märt Roosaare
Date of birth:	May 16, 1987
Contact:	+ 372 737 4054
E-mail:	mart.roosaare@ut.ee

Education:

University of Tartu, PhD student in Gene technology, 2014-…
University of Tartu, MSc in Gene technology, 2014.
University of Tartu, BSc in Gene technology, 2011.

Professional employment:

2012–2018 University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology, Programmer

Publications:

Kõiv, V.; Roosaare, M.; Vedler, E; Kivistik., P A; Toppi, K.; Schryer, DW.; Remm, M.; Tenson, T; Mäe, A. (2015). Microbial population dynamics in response to Pectobacterium atrosepticum infection in potato tubers. Scientific Reports, 5 (11606), 1–18.10.1038/srep11606.

Roosaare, M.; Vaher, M.; Kaplinski, L.; Möls, M.; Andreson, R.; Lepamets, M.; Kõressaar, T.; Naaber, P.; Kõljalg, S.; Remm, M. (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ 5:e3353; DOI 10.7717/peerj.3353.

Roosaare, M.; Puustusmaa, M.; Möls, M.; Vaher, M.; Remm, M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ 6:e4588; DOI 10.7717/peerj.4588.

Supervised dissertations:

Mihkel Vaher, Master's Degree, 2016, (sup) Märt Roosaare, Identifying bacterial strains from unassembled sequencing reads using fixed length oligomers, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

ELULOOKIRJELDUS

Nimi:	Märt Roosaare
Sünniaeg:	16. mai 1987
Aadress:	Bioinformaatika õppetool, Molekulaar- ja rakubioloogia instituut, Riia 23B, 51010, Tartu
Telefon	737 4054
E-post:	mart.roosaare@ut.ee

Haridus:

Tartu Ülikool, doktorant geenitehnoloogia erialal, 2014-...
Tartu Ülikool, magistrikraad geenitehnoloogia erialal, 2014.
Tartu Ülikool, bakalaureusekraad geenitehnoloogia erialal, 2011.

Teenistuskäik:

2012–2018 Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut, programmeerija

Teaduspublikatsioonid:

Kõiv, V.; Roosaare, M.; Vedler, E; Kivistik., P A; Toppi, K.; Schryer, DW.; Remm, M.; Tenson, T; Mäe, A. (2015). Microbial population dynamics in response to Pectobacterium atrosepticum infection in potato tubers. Scientific Reports, 5 (11606), 1–18.10.1038/srep11606.

Roosaare, M.; Vaher, M.; Kaplinski, L.; Möls, M.; Andreson, R.; Lepamets, M.; Kõressaar, T.; Naaber, P.; Kõljalg, S.; Remm, M. (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ 5:e3353; DOI 10.7717/peerj.3353.

Roosaare, M.; Puustusmaa, M.; Möls, M.; Vaher, M.; Remm, M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ 6:e4588; DOI 10.7717/peerj.4588.

Juhendatud väitekirjad:

Mihkel Vaher, magistrikraad, 2016, (juh) Märt Roosaare, Bakteritüvede tuvastamine sekveneerimise toorlugemitest kindla pikkusega oligomeeride abil, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan**. Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.
41. **Sulev Ingerpuu**. Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.

62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.
63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.

83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.
84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli**. The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.

104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.
106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Kaasik, Krista**. Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš**. The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman**. Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja**. Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras**. Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu**. Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar**. The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper**. Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves**. Stress-induced transposition of Tn*4652* in *Pseudomonas Putida*. Tartu, 2006, 120 p.
118. **Silja Kuusk**. Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa**. Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht**. Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk**. Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann**. Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers**. Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli**. Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.

125. **Kai Rünk**. Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.
126. **Aveliina Helm**. Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007, 89 p.
127. **Leho Tedersoo**. Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007, 233 p.
128. **Marko Mägi**. The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007, 135 p.
129. **Valeria Lulla**. Replication strategies and applications of Semliki Forest virus. Tartu, 2007, 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007, 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007, 171 p.
132. **Tatjana Krama**. Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007, 112 p.
133. **Signe Saumaa**. The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida*. Tartu, 2007, 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007, 96 p.
135. **Priit Kilgas**. Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007, 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007, 95 p.
137. **Kärt Padari**. Protein transduction mechanisms of transportans. Tartu, 2008, 128 p.
138. **Siiri-Lii Sandre**. Selective forces on larval colouration in a moth. Tartu, 2008, 125 p.
139. **Ülle Jõgar**. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p.
140. **Lauri Laanisto**. Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008, 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008, 105 p.
142. **Birgot Paavel**. Bio-optical properties of turbid lakes. Tartu, 2008, 175 p.
143. **Kaire Torn**. Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg**. Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd**. Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.

146. **Lauri Saag**. Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu**. Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm**. Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks**. Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu**. Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš**. Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko**. Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast**. Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats**. Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova**. The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik**. Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver**. Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja**. Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast**. Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak**. Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel**. Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe**. Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe**. Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.

165. **Liisa Metsamaa**. Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.
166. **Pille Säälik**. The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil**. Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik**. Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark**. Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap**. Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan**. Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe**. Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi**. Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson**. Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts**. Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis**. Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov**. Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster**. Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap**. Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar**. Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül**. Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk**. Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü**. Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla**. Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.

185. **Gyaneshwer Chaubey**. The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.

186. **Katrin Kepp**. Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.

187. **Virve Sõber**. The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.

188. **Kersti Kangro**. The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.

189. **Joachim M. Gerhold**. Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.

190. **Helen Tammert**. Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.

191. **Elle Rajandu**. Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.

192. **Paula Ann Kivistik**. ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.

193. **Siim Sõber**. Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.

194. **Kalle Kipper**. Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.

195. **Triinu Siibak**. Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.

196. **Tambet Tõnissoo**. Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.

197. **Helin Räägel**. Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.

198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.

199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.

200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.

201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.

202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.

203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.

205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.

224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.
225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.

245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.
246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.

263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.
265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p

282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.
284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.
285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.

301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.
303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.

321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.
324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.
325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik.** Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak.** Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak.** Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar.** Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo.** Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova.** Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare.** The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.