MARTTI VASAR

Developing a bioinformatics pipeline gDAT
to analyse arbuscular mycorrhizal fungal
communities using sequence data
from different marker regions

MARTTI VASAR

Developing a bioinformatics pipeline gDAT to analyse arbuscular mycorrhizal fungal communities using sequence data from different marker regions

Department of Botany, Institute of Ecology and Earth Sciences, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation was accepted for the commencement of the degree of *Doctor philosophiae* in plant ecology at the University of Tartu on April 6th, 2020 by the Scientific Council of the Institute of Ecology and Earth Sciences, University of Tartu.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The thesis is based on the following papers, which are referred to in the text by the relevant Roman numerals:

I. Thiéry O, **Vasar M**, Jairus T, Davison J, Roux C, Kivistik PA, Metspalu A, Milani L, Saks Ü, Moora M, Zobel M, Öpik M. 2016. Sequence variation in nuclear ribosomal small subunit, internal transcribed spacer and large subunit regions of Rhizophagus irregularis and Gigaspora margarita is high and isolate-dependent. *Molecular Ecology*, 12:2816–2832.

II. **Vasar M**, Andreson R, Davison J, Jairus T, Moora M, Remm M, Young JP, Zobel M, Öpik M. 2017. Increased sequencing depth does not increase captured diversity of arbuscular mycorrhizal fungi. *Mycorrhiza*, 8:761–773.

III. Lekberg Y, **Vasar M**, Bullington LS, Sepp SK, Antunes PM, Bunn R, Larkin BG, Öpik M. 2018. More bang for the buck? Can arbuscular mycorrhizal fungal communities be characterized adequately alongside other fungi using general fungal primers? *New Phytologist*, 4:971–976.

IV. **Vasar M**, Davison J, Neuenkamp L, Sepp SK, Young JP, Zobel M, Öpik M. User-friendly bioinformatics pipeline gDAT (graphical downstream analysis tool) for analysing rDNA sequences. Manuscript

Published papers are reproduced with the permission of the publishers.

Author's contribution to each publication

	I	**II**	**III**	**IV**
Idea and design	−	+	−	+
Software development	−	+	−	+
Data analysis	+	+	+	+
Writing	+	+	+	+

List of abbreviations

AM – arbuscular mycorrhiza
BLAST – Basic Local Alignment Search Tool
DNA – deoxyribonucleic acid
HTS – high-throughput sequencing
INSDC – International Nucleotide Sequence Database Collaboration
ITS – Internal Transcribed Spacer
LSU – large subunit
OTU – operational taxonomic unit
PCR – polymerase chain reaction
rDNA – ribosomal deoxyribonucleic acid
RNA – ribonucleic acid
SSU – small subunit
VT – virtual taxon

List of terms and definitions

barcoding – a method of species identification from a sample containing single organism using a short section of DNA from a specific gene or genes.

BLAST – Basic Local Alignment Search Tool, an algorithm to align input sequences with reference sequences by maximizing their local score, outputting alignment length, identities, gaps and mismatches.

chimera (chimeric sequence) – artifact sequence containing DNA from multiple organisms.

Maarj*AM* database – manually curated online database containing DNA sequences of AM fungi from the SSU, ITS and LSU rRNA gene region that originate from ecological studies based on environmental samples or taxonomic investigations based on cultured fungi. SSU marker gene data is used to define phylogenetically determined species estimates called VT (virtual taxon).

metabarcoding – a method of species identification from a sample containing DNA from more than one organism.

sequence clustering – process of grouping sequences into molecular operational taxonomic units (OTU), utilizing specific algorithms (heuristic approach or single-linkage) and thresholds of similarity or dissimilarity (cutoff).

sequence identification – identifying a sequence against a model or a reference sequence database of taxonomic information.

UNITE database – general fungal sequence database containing DNA sequences from the ITS marker region, where sequences are delimited into species hypothesis (SH) based on clustering and curation by expert taxonomists.

virtual taxon – a custom taxonomic unit of AM fungi, which has been delimited based on the SSU rRNA gene sequences and implemented in the Maarj*AM* database. These are phylogenetically defined taxonomic units delimited mostly at 97–99% sequence similarity.

1. INTRODUCTION

Arbuscular mycorrhizal (AM) fungi (Phylum Mucoromycota, Subphylum Glomeromycotina, Spatafora et al. 2016) are widespread soil microorganisms that are obligate symbionts of the large majority of land plants and play essential roles in ecosystems by improving nutrient uptake and resistance to biotic and abiotic stress in their hosts, and by influencing soil structure (Smith and Read 2008; Tedersoo, Bahram and Zobel 2020). AM fungal diversity is related to important ecosystem properties and processes, including plant diversity, primary productivity and plant invasions (van der Heijden, Bardgett and van Straalen 2008; Moora et al. 2011). AM fungi play critical roles in carbon and nutrient cycles around the globe (van der Heijden et al. 2015; Treseder 2016). To understand the mechanisms underlying these properties, there is increasing interest in studying communities of AM fungi in natural conditions, and considerable effort has been made to study AM fungal diversity and how it relates to ecosystem functioning (van der Heijden et al. 2015; Treseder 2016).

Traditionally, AM fungal communities have been studied by using trap-cultures to capture and extract spores from the soil, and identifying collected spores on the basis of morphological traits (Smith and Read 2008; Öpik et al. 2014). Detecting and identifying AM fungi in natural systems is not straightforward since the organisms are microscopic and often not culturable in controlled conditions together with plant hosts (Smith and Read 2008). AM fungi are considered asexual clonal organisms making it difficult to define them using a biological species concept (but see Corradi and Brachmann 2017). Identification of spores is also hampered by a shortage of morphological characteristics that distinguish closely related taxa, while it is also possible that some AM fungal species might not form spores at all. The difficulty of distinguishing taxa both at morphological and genetic levels hampers assessments of the diversity of AM fungal communities in natural environments and obstructs understanding of the processes influencing AM fungal and plant interactions.

DNA sequence-based molecular methods have provided the means to overcome indistinguishable morphological characteristics and poor culturability, allowing samples from natural conditions to be studied (Öpik and Davison 2016). DNA sequence-based fungal identification does not necessarily require taxonomic expertise in specific fungal groups any longer. Rapid advances have made sequence-based identification a widespread and cost-efficient method for assigning taxonomic identities to culturable and unculturable organisms alike, improving the local and global scale precision of biodiversity estimates (Lindahl et al. 2013; Hart et al. 2015; Taylor, Helgason and Öpik 2017). During recent decades, multiple technologies for acquiring sequence data from natural samples have been developed, with cloning followed by Sanger sequencing being the first method that stimulated sequencing of soil microbes (e.g., Öpik et al. 2008). Newer DNA sequencing technologies, called high-throughput sequencing (HTS), have overcome the constraints of the cloning step that was needed in the case of

Sanger-sequencing for separating different reads in a sample and can rapidly generate large numbers of reads of sufficient length suitable for metabarcoding and a deep assessment of natural AM fungal diversity (Taberlet et al. 2012; Cui et al. 2016; Johansen et al. 2016; Xu et al. 2016; Orchard et al. 2017; Wang, White and Li 2016). 454-pyrosequencing was the next major method developed for analysing environmental samples, providing up to 600 base pair length sequences (Shokralla et al. 2012), which is suitable for a large range of primer pairs used in various marker gene regions. By now, 454-squencing has been largely replaced by Illumina-sequencing, which generates many times more sequences, but with reduced sequence length, providing paired-end reads of 300 bases that can be combined almost up to 600 bases long reads depenent on the overlap region size. HTS provides sufficient sequence length to cover most popular marker regions. Sequence counts produced by HTS are orders of magnitude higher resulting in higher feasible sequence numbers per sample that has increased probability of detecting rare taxa as compared to Sanger-sequencing, improving our knowledge about global biodiversity distribution patterns and potential underlying drivers of biodiversity (e.g., Öpik et al. 2009; Davison et al. 2015). The technological advance has also returned research focus to analyse AM fungi in soil samples, compared to the initial interest in actively symbiosis forming root colonizing AM fungi. Sanger-sequencing was generally applied to investigations of root or spore samples, which also tended to have high relative concentrations of AM fungal DNA, while full extraradical AM fungal communities (including hyphae and spores) were often not considered (Lumini et al. 2010). Targeting only root data may not capture the full assemblage of AM fungi present in soil as roots contain only subset of the taxa present in the soil, as was further clarified with the help of HTS (Saks et al. 2014; Varela-Cervero et al. 2015; Sepp et al. 2018). More broadly speaking, metabarcoding of soil samples can be used to study a range of other soil microbial groups from the same DNA extracts, opening new research directions such as analyzing organisms' co-variation (e.g., Noreika et al. 2019).

Some standard approaches have been developed for sequencing the DNA of AM fungal taxa present in environmental samples (Öpik et al. 2010; Kivlin, Hawkes and Treseder 2011; Öpik et al. 2013; Davison et al. 2015). Nuclear ribosomal operon regions such as the small subunit (SSU) and large subunit (LSU) rRNA genes and the internal transcribed spacer (ITS) region are the most commonly used markers for approximately species level identification of AM fungi in community surveys (Kohout et al. 2014; Öpik et al. 2014; Hart et al. 2015). DNA sequence-based species identification is possible if variation in a marker sequence is lower within species than between species, i.e., if the marker possesses a barcode gap (Meyer and Paulay 2005). Validation of the barcode region, finding suitable primers and finding or compiling an appropriate reference database is crucial to achieve correct species identification of any organism. A lot of work has been done to sequence culturable fungi and morphologically distinct species with various primers and the data have been made available through multiple reference databases (Maarj*AM* database, Öpik et al. 2010; UNITE database, Abarenkov et al. 2010; PHYMYCO-DB, Mahé et al. 2012). Some primers

might not amplify all AM fungal families with equal efficiency, and changes to the primer design have to be made to improve the capture rate (Ihrmark et al. 2012; Kohout et al. 2014). Whole genome sequence data can be used to identify potential new primer sites (Wang and Qian 2009; Lin et al. 2014).

DNA sequence-based identification of organisms requires availability of a reference database for sequence comparisons. Reference databases should ideally feature a large taxonomic coverage and include well-annotated sequences of good quality and known origin that have been identified by experts. In order to uncover AM fungal diversity in more detail than it is possible by using morphological characters, a custom unit of molecular diversity – the virtual taxon (VT) – has been proposed and implemented in the Maarj*AM* database (Öpik et al. 2010). The Maarj*AM* database contains Glomeromycotina DNA sequence data that originate from published ecological studies based on environmental samples or taxonomic investigations of cultured fungi. VT are phylogenetically defined taxonomic units corresponding to 97–99% sequence identity. The Maarj*AM* database covers reference sequences from SSU, ITS and LSU marker regions; though VT are only constructed for the SSU rDNA region.

Advances in molecular identification tools have greatly improved our understanding of AM fungal communities globally (Davison et al. 2015). There are many tools that help to analyse biological data allowing clustering, chimera checking and matching sequences against a reference database (Caporaso et al. 2010; Magoč and Salzberg 2011; Rognes et al. 2016; Camacho et al. 2009), but these can be challenging for users without bioinformatics training. Multiple general-purpose pipelines have been produced to allow more easy-to-use means for HTS analyses (Caporaso et al. 2010; Anslan et al. 2017), which helps to reduce knowledge gap about available bioinformatics workflows and providing methods to handle different file formats. New sequencing technologies, methods, protocols and markers can hinder these applications as the pipelines are not always updated to support the latest sequencing methods. Lack of bioinformatics expertise may pose a serious bottleneck for barcoding studies and potentially result in incorrect conclusions being drawn from sequence data.

The outcome of this thesis is the bioinformatics pipeline gDAT (graphical downstream analysis tool). The pipeline aims to help ecologists to perform meta-barcoding studies based on HTS data of ecological samples, using various marker regions and genes. The pipeline combines different third-party bioinformatics software into easy-to-use software and includes the well-annotated Maarj*AM* database as a reference dataset allowing the opportunity to implement the VT molecular taxonomy. Although the pipeline is optimized for use for AM fungal identification with the Maarj*AM* database using the SSU rDNA marker gene, it supports the use of other reference databases and marker regions and hence can be used to identify sequences of any organism. The SSU rDNA marker gene has proven to be the most widespread region for analyzing AM fungi, and it exhibits a highly variable region flanked by invariable regions and a concise read length, making it suitable for sequence alignment and phylogenetic analyses (**I**). Furthermore, primer pairs for the SSU gene region exist that can amplify most

known AM fungal families and can capture various soil organisms besides AM fungi, allowing more in-depth analysis of microorganisms occurring in plant roots or soil (Öpik et al. 2010). This pipeline is optimized to run on commodity hardware. The pipeline is built using Python scripting language, supporting both version 2.6+ and version 3+ to cover a wide variety of setups and configurations.

The thesis has the following aims:
1) Analysing the suitability of different marker regions for identifying AM fungi, including measures of inter- and intraspecific variation (**I**, **III**);
2) Comparing 454-sequencing with newer Illumina-sequencing technology in terms of their detection of abundant and rare taxa, and community composition changes (**II**);
3) Comparing AM fungal richness and community composition on the basis of the two most commonly used marker regions in fungal ecology: ITS and SSU and general fungal vs AM fungal specific primers, respectively (**I**, **III**);
4) Developing a pipeline for analysis of (AM) fungal sequences using HTS and developing a graphical user interface to provide easy-to-use software for ecologists (**II**, **IV**).

2. MATERIALS AND METHODS

2.1. Study design, sites and data collection

Paper **I** assessed the suitability of marker regions by comparing intra- and inter-specific variation among two Glomeromycotan species: *Rhizophagus irregularis* and *Gigaspora margarita*. Three marker regions were amplified separately from the same single spore DNA extracts and each amplicon was cloned: about 1760 bp of the SSU, the full ITS and about 2010 bp of the LSU region, spanning the ribosomal DNA region (45S) of *ca* 4370 bp. Up to 5 spores from two isolates for each species were collected and sequenced using a Sanger-sequencing platform. Variation within spore, within isolate, within species and between species was measured for each region and for each species. Methods for separating spores, identifying spores and DNA extraction are described in paper **I**.

Paper **II** studied the difference between two sequencing technologies: 454 and Illumina MiSeq. AM fungal amplicons of the nuclear SSU rRNA gene were generated with primers NS31 and AML2 (Simon, Lalonde and Bruns 1992; Lee, Lee and Young 2008; Table 1). However, at the time the analyses were conducted, Illumina MiSeq only produced 2×250 bp paired-end reads, which was not sufficient to cover SSU gene region using NS31 and AML2 primers. To overcome this limitation a tagmentation based method was used that permit sequencing longer DNA fragments. Tagmentation incorporates universal tags into DNA fragments of user-selected size via transposon cleavage (Lee et al. 2016), allowing longer fragments, such as the full ribosomal small subunit RNA genes of bacteria, to be sequenced (Burke and Darling 2016). Selection of different plant individuals and plant root sampling was conducted in the Järvselja forest reserve in south-eastern Estonia (58° 17.916' N, 27° 15.744' E), as described in Saks et al. (2014). Samples were collected from two plots of 30 x 30 m, about 50 m apart, in June 2009. Two AM plant species that were abundant in the plots were sampled for this study: *Convallaria majalis* L. and *Rubus saxatilis* L. Three random individuals of each plant species per plot were excavated (a total of 12 samples). In addition to comparing 454-sequencing with Illumina-sequencing, paper **II** also assessed whether AM fungal data could be recovered from metagenomic DNA extracted from plant roots which avoids the PCR amplification step and decreases the chimeric read formation.

Primers have been constantly developed and upgraded to match current sequencing technologies and to improve the capture rate of large groups of organisms and families. Paper **III** analysed the ability of general fungal ITS and and AM fungal specific SSU primers to capture AM fungal community composition. For ITS improved general fungal primers were used to allow capture of AM fungal sequences (Ihrmark et al. 2012; Kohout et al. 2014; White et al. 1990). However, SSU region has widely used primers designed for capturing specifically AM fungi (Lee, Lee and Young 2008; Dumbrell et al. 2011), which can be used to check if AM fungal sequences were present in the samples and if the

composition for different treatments used in the study have similar patterns with the AM fungal sequences captured by the ITS region. Two experiments were set up using root samples: (i) to measure soil moisture effect in a greenhouse experiment; (ii) and to measure seasonality using data from a field experiment by measuring AM distribution in April, July and September in 2016. The greenhouse experiment involved spotted knapweed (*Centaurea stoebe*) and mountain brome (*Bromus marginatus*) grown for three months in spotted knapweed rhizosphere soil under either high (~30%) or low (~10%) soil moisture prior to harvesting roots for analyses. The field survey consisted of a seasonal sampling of showy milkweed (*Asclepias speciosa*) roots from intermountain grassland in Montana, USA. Here we use significant shifts observed in the SSU datasets due to soil moisture and season to ask if the AM fungi amplified using general fungal ITS primers and the same DNA templates report similar shifts in composition. Portions of the SSU and the ITS regions were amplified using the primer pairs WANDA-AML2 (Lee, Lee and Young 2008; Dumbrell et al. 2011) and fITS7:fITS7o/ITS4 (Ihrmark et al. 2012; Kohout et al. 2014; White et al. 1990), respectively, where the fITS7o is a modification of fITS7 with improved detection of all AM fungal families. For SSU, the NS31 primer was replaced by the WANDA primer to improve the overlap region length, as the WANDA primer is located after NS31 primer making the amplified region shorter and thus increasing the overlap region of read 1 and read 2. We used a semi-nested PCR for both primer pairs where barcodes and tags were added in the second PCR step. PCR products were cleaned, quantified and pooled within studies at equimolar concentrations prior to Illumina MiSeq sequencing using 2×300 bp paired-end reads at the University of Idaho (www.ibest.uidaho.edu).

The final paper (**IV**) integrates knowledge from the previous studies to develop a general-purpose pipeline for studying and analysing AM fungal data from various rDNA regions. This tool uses Python language to offer accessibility for large groups of users as it is pre-installed or can be easily installed on various operating systems. The developed pipeline helps scientists who have minimal knowledge about programming and the command line to use a graphical interface with provided manual including step-by-step instructions and tooltips to analyse their data on the go. This shifts scientists' efforts and focus from analysing the data to making and drawing ecological conclusions from the results. In order to validate the pipeline performance and capabilities, an unpublished sequencing data set was used (Sepp et al. unpublished). There were a total of 36 soil samples collected from Laelatu wooded meadow located near Virtsu village on the western coast of Estonia (58° 35.10' N, 23° 34.03' E). Sampling was conducted in four plots of 30×30 m in the central part of the meadow. Two plots were located in open parts of the meadow without any trees or shrubs (except single 1-year old saplings) within a radius of 30 m. The remaining two plots were located under small woody groves, mainly of *Quercus robur* and *Corylus avellana*, standing within a grassland matrix. AM fungal sequences were amplified from soil DNA extracts using the AM fungal specific primers for the SSU ribosomal RNA gene V4 region: WANDA (Dumbrell et al. 2011) and AML2

(Lee, Lee and Young 2008; Table 1). The total fungal community was identified by sequencing the ITS2 region with degenerate primer pair fITS7 (forward; Ihrmark et al. 2012) and ITS4 (reverse primer; White et al. 1990). Degenerate primers were used to allow for maximum detection of various fungal taxa. The total eukaryotic community was identified by sequencing the SSU rRNA gene V4 region with primers F574 and R952 (Hadziavdic et al. 2014).

Table 1. Primers used in papers **I**, **II**, **III** and **IV**.

Marker	Primer	Sequence 5' - 3'	Reference	Paper
SSU	NS31	TTGGAGGGCAAGTCTGGTGCC	Simon, Lalonde and Bruns 1992	**I, II**
SSU	WANDA	CAGCCGCGGTAATTCCAGCT	Dumbrell et al. 2011	**III, IV**
SSU	F574	GCGGTAATTCCAGCTCCAA	Hadziavdic et al. 2014	**IV**
SSU	R952	TTGGCAAATGCTTTCGC	Hadziavdic et al. 2014	**IV**
SSU	AML2	GAACCCAAACACTTTGGTTTCC	Lee, Lee and Young 2008	**I, III, IV**
ITS	fITS7	GTGARTCATCGAATCTTTG	Ihrmark et al. 2012	**III, IV**
ITS	fITS7o	GTGAATCATCRAATYTTTG	Kohout et al. 2014	**III**
ITS	ITS4	TCCTCCGCTTATTGATATGC	White et al. 1990	**III, IV**
LSU	Glo454	TGAAAGGGAAACGATTGAAGT	Lekberg et al. 2012	**I**
LSU	NDL22	TGGTCCGTGTTTCAAGAC	van Tuinen et al. 1998	**I**
LSU	FLR3	TTGAAAGGGAAACGATTGAAG	Gollotte, van Tuinen and Atkinson 2004	**I**
LSU	FLR4	TACGTCAACATCCTTAACGAA	Gollotte, van Tuinen and Atkinson 2004	**I**

2.2. gDAT pipeline

HTS of multiple organisms at the same time from the same samples (metabarcoding, see Taberlet et al. 2012) has become a routine and cost-effective method for analysis of microbial communities in environmental samples. However, careful data treatment is required to identify potential errors in HTS data, and the large volume of data generated by HTS requires in-house experience with command-line tools for downstream analysis. The developed bioinformatic pipeline gDAT has a graphical user interface for analyzing HTS data and has been designed to work on a commodity computer with limited CPU and memory, though multiple cores allow some of the programs to run in parallel, reducing overall runtime significantly. The pipeline is written in Python language supporting both 2.6+ and 3+ versions allowing it to be used with variety of operating systems and computers. Multiple third-party software (Table 2) has been incorporated into the pipeline to allow combination of paired-end sequences (FLASh), sequence clustering (VSEARCH), chimera checking (VSEARCH) and sequence identification (BLAST). gDAT, which is described in paper **IV,** can be downloaded from GitHub (https://github.com/ut-planteco/gDAT); and an earlier version of the pipeline, which was used in papers **II** and **III,** can be downloaded from GitHub (https://github.com/ut-planteco/ssu-pipeline).

Table 2. List of main and third-party applications incorporated into the pipeline to conduct DNA-based analyses.

Analysis step	Program	Version	Reference
Combine paired-end sequences	FLASh	1.2.11	Magoč and Salzberg 2011
Chimera filtering	vsearch	1.11.1	Rognes et al. 2016
Clustering	vsearch	1.11.1	Rognes et al. 2016
Taxonomy assignment	BLAST+	2.5.0+	Camacho et al. 2009
Pipeline	gDAT	1.0	github.com/ut-planteco/gDAT
Earlier version of the pipeline	ssu-pipeline	1.0	github.com/ut-planteco/ssu-pipeline

2.3. Data analyses

Differences between the methods used for AM fungal sequence identification revealed in the studies (**I**, **II**, **III**, **IV**) reflect the state of the art in terms of available sequencing platforms, primers and software used and the focus of research questions at the respective times of publishing.

Forward and reverse sequences retrieved from individual clones in paper **I** were overlapped and assembled into contigs using CONSED v2.1 (Gordon et al. 2004). Chimeras were checked with UCHIME using reference database mode (Maarj*AM* for SSU and LSU, UNITE for ITS; Öpik et al. 2010; Nilsson et al. 2019). ITS marker amplicons (containing partially flanking SSU and LSU regions) were trimmed to contain only the ITS region. The SSU and LSU rRNA gene amplicons were used in a full length as they did not contain flanking markers. To compare environmental sequence identification procedures with sequences of known identity, the Sanger sequences from paper **I** were processed using QIIME pipeline (Caporaso et al. 2010). Sequences were trimmed to the most commonly used amplicons/regions used in AM fungal community surveys: the NS31-AML2 amplicon for SSU, the ITS1-5.8S-ITS2 region for ITS and the Glo454-NDL22 amplicon for LSU (Table 1). For SSU and LSU, sequences from the Maarj*AM* database (Öpik et al. 2010; status October 2015) were used and for ITS the UNITE database was used (Abarenkov et al. 2010; status October 2015). The following QIIME pipeline data handling steps were applied to Sanger sequences: (i) OTUs (species proxies) picked with UCLUST using the default method that generates *de novo* OTUs based on 97% identity threshold; (ii) picking representative sequences for each OTU; (iii) assigning taxonomy to the representative sequence set for each OTU. OTUs were picked using a range of cut-off thresholds from 95% to 99%. To test intra- and interspecific variation, the sequence set was BLASTed against itself to record similarity values within spores, isolates and species, and between species.

In paper **II** reads were subjected to the bioinformatics procedures described in Öpik et al. (2013) and Saks et al. (2014). In short, 454-sequencing reads were retained if they carried the correct barcode (8 bp) and forward primer sequence (21 bp) and were ≥170 bp long (excluding barcode and primer sequence). Barcode and primer sequences were stripped from reads. Sequences longer than 520 bases were trimmed to exclude reverse primer information. Illumina amplicon reads were quality-filtered by removing sequences exhibiting Nextera adapter contamination and where average quality was <30 (maximum 41; Cock et al. 2010). Adapter fragments were revealed at random locations dependent on insert size, which was sometimes smaller than MiSeq read length, resulting in palindrome reads where adapter read-through occurs (overhang sequences; Bolger, Lohse and Usadel 2014). After removal of sequences containing adapter and those with low average quality, reads without the respective paired read were omitted, as they could not be combined. Remaining reads were combined using FLASh software with the default parameters (minimum overlap = 10 bp; maximum mismatch density in the overlap segment = 0.25). 454 and Illumina reads containing potential chimeras were detected and removed using USEARCH in reference database mode (Maarj*AM* database, status 2017) using the default settings (Edgar et al. 2011). Chimera free reads were identified against the most variable part of the AM fungal SSU sequences between primers NS31 and AML2 in the Maarj*AM* database (Figure 1) using BLAST by trimming input sequences between positions 70 and 300 bases after the NS31 primer. Those reads that BLAST did not match

against the Maarj*AM* database (no-hits) were investigated by conducting a further BLAST search against the INSDC non-redundant nucleotide database (status October 2016; Benson et al. 2013). Analysis of metagenomic Illumina reads followed the same approach as Illumina amplicon-based reads but included a further filtering step based on GC% content between 20 to 42% following the previously published AM fungal genome data (Tisserant et al. 2013; Lin et al. 2014). Remaining sequences were assembled into contigs with Newbler (v2.6; the 454 Life Science de novo assembler) using the default parameters (minimum identity 96%, minimum length 40 bases and parameter *–large* to speed up assembly and reduce memory footprint). Potential genes from assembled contigs were predicted and translated into proteins using GeneMark-ES (v4.33; Ter-Hovhannisyan et al. 2008) with the prebuilt Hidden Markov Model for eukaryotes. Putative proteins were subjected to a BLAST search against the INSDC non-redundant protein database (status December 2016) using protein BLAST with 40% identity (Rost 1999) and 50% alignment filter of the shorter read (query or database sequence) giving rough estimate of the organisms found in samples.
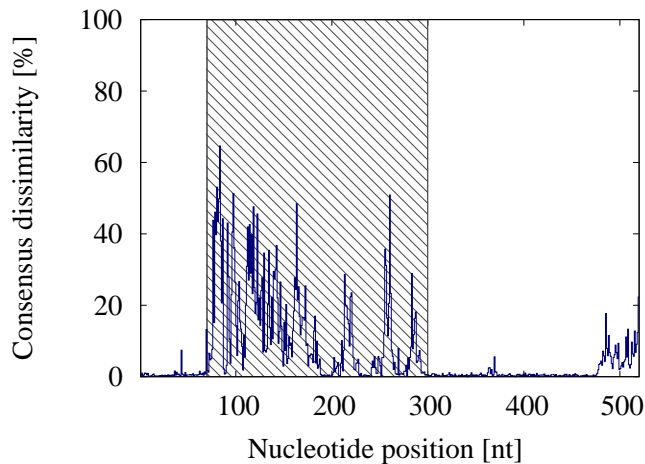

Figure 1. Consensus dissimilarity calculated for SSU rRNA gene region using reference sequences from the Maarj*AM* database located between NS31-AML2 primers (**II**). Shaded area displays the variable region of the amplicon. [Adapted by permission from Copyright Clearance Center: Springer Nature Mycorrhiza "Increased sequencing depth does not increase captured diversity of arbuscular mycorrhizal fungi", Vasar et al. 2017, Copyright 2017]

Paper **III** tested the performance of ITS to capture AM fungal diversity using an improved general fungal primer and examined community composition differences compared to AM fungal specific SSU amplicon sequencing under different treatments, using the custom pipeline from paper **II** for cleaning and analysing the data. Illumina MiSeq paired-end raw reads (2 × 300 bp) were demultiplexed into samples using the correct barcode pairs (8 bp + 8 bp). Demultiplexed samples

were cleaned by checking for the correct primer pairs and average quality of at least 30 for both sequences. Barcode and primer sequences were excluded from reads. As the ITS2 region of AM fungi is shorter than the SSU region, and to improve overlap success rate, forward reads were trimmed to 220 bases and reverse reads to 180 bases to remove low quality regions in the sequences. Retained sequences were combined using FLASh (v.1.2.10; Magoč and Salzberg 2011) with default parameters (minimum overlap 10 bp and alignment identity ≥75%). Combined reads were checked for chimeras using USEARCH (v7.0.1090; Edgar et al. 2011) using database mode with default parameters using the Maarj*AM* database (status October 2017; SSU partition contained 469 VT, ITS partition contained 391 OTUs). As ITS sequences in the Maarj*AM* database are not organized into VT, we clustered ITS reference sequences into OTUs using BLASTclust (legacy BLAST v2.2.26; Altschul et al. 1990) with 97% sequence identity, for comparison with SSU VT. After removal of chimeric sequences, the remaining sequences were identified using a BLAST (v2.5.0+; Camacho et al. 2009) search against the Maarj*AM* database using criteria for a match as in Davison et al. (2012): sequence similarity ≥97%; an alignment length ≥95% of the shorter of the query read and reference read; and a BLAST e-value <1e-50.

In paper **IV** the custom pipeline used in papers **II** and **III** was consolidated and a graphical interface and extensive manual added to support users with analysis of HTS data. To test the capability of the pipeline, metabarcoding data from a pilot study from Sepp et al. (unpublished) was used to test multiple marker regions and different reference databases performance. AM fungal SSU sequences from soil samples were amplified using WANDA-AML2 (Dumbrell et al. 2011; Lee, Lee and Young 2008, Table 1) primers using 2 × 300 bp Illumina MiSeq v3. The same soil samples were amplified using improved general fungal ITS primers fITS7-ITS4 (Ihrmark et al. 2012; White et al. 1990) and general eukaryotic primers F574-R952 (Hadziavdic et al. 2014) to investigate organism distribution and diversity in the soil using 2 × 250 bp Illumina MiSeq v3. General eukaryotic sequences were used to test how many AM fungal sequences can be captured by using BLAST against the Maarj*AM* database (status July 2019) compared to AM fungal sequences. ITS sequences were identified with BLAST against the UNITE database (status February 2019). Sequences were cleaned using barcode and primer checking, allowing 1 nucleotide mismatch for each forward and reverse pair. An average quality threshold of 30 was used for both sequences. ITS sequences were trimmed as in paper **III** to 220 bp and 180 bp for forward and reverse reads respectively to remove low quality regions and improve the combination rate. Reads were combined using FLASh with 10 bp minimum overlap length and 75% identity for overlap region. Combined sequences were searched for chimeras using Maarj*AM* for the SSU region and UNITE for the ITS region. BLAST searches were conducted against Maarj*AM* for the SSU region and against UNITE for the ITS region using e-value <1e-50. The results were generated into pivot tables from the BLAST output using 97% identity and 95% alignment thresholds. No-hits were subjected to BLAST against INSDC non-redundant nucleotide database to identify non-AM fungal organisms

present in the samples. Hits against INSDC were summarized using 95% identity and 90% alignment thresholds and using a common taxonomy, which was generated by analysing the 500 best hits and using the lowest common taxonomic rank for which a common identity occurred in at least 51% of the hits.

2.4. Statistical analyses

In order to estimate the ability of the sequencing effort (obtained sequences per sample) to detect all sequence variants present in paper **I**, rarefaction curves (sequence variant accumulation curves) were calculated using the function rarefy from the R package vegan (Oksanen et al. 2012). Nucleotide sequence divergence (mean evolutionary distance, MED) within and between spores was calculated pairwise for each ribosomal gene region based on the Kimura 2-parameter model with pairwise deletion of gaps using MEGA v.4.0.2 (Tamura et al. 2007). Nucleotide diversity (π) within individual marker regions was calculated as the average number of differences per base between sequences compared pairwise and plotted for 10-base windows using dnaSP v.5.10.01 (Librado and Rozas 2009).

It is common practice to exclude singleton taxa (i.e. taxa represented by a single DNA read) or rare taxa (i.e. taxa below a specified abundance) prior to conducting diversity analyses using DNA based community data. This is because singleton or rare taxa may represent sequencing artefacts (Tedersoo et al. 2010) or may be of minor interest if a description of dominant community patterns is intended. In paper **II** we removed singleton taxa from 454 sequencing results, where each singleton hit represented 0.01% of all the hits. Using the same proportion, we removed rare taxa from Illumina reads when they represented less than 0.01% of total hits. Sampling efficacy was assessed using individual-based rarefaction (function rarefy from R package vegan; Oksanen et al. 2016). Richness estimates were regressed against read number within each sequence set and against each other between sets. The relative abundances of different VT in the 454 and Illumina sequence sets (samples pooled within each set) were compared using Pearson's correlation. Pearson's correlation was also used to test the relationships between sample weight and recorded VT richness and sequence depth.

In paper **III** VT and OTU accumulation curves were generated for each sample using QIIME (v1.9.1; Caporaso et al. 2010). Sample standardization to common sequencing depth was conducted within QIIME. Bray-Curtis distance matrices of Hellinger transformed data were generated for each rarefied OTU and VT table and used to detect differences in community composition between soil moisture treatments and across seasons using an analysis of similarities (ANOSIM).

3. RESULTS AND DISCUSSION

Studying AM fungal species in soils is not straightforward as the fungi are often unculturable (Tedersoo, Bahram and Zobel 2020), but emerging sequencing technologies have made it possible to analyse soil organisms in natural conditions (Öpik et al. 2010). The new sequencing technologies have generated an enormous amount of environmental DNA sequences, providing new insights but also challenges in the field of molecular ecology. One challenge is the allocation of sequences to taxonomic units that can provide meaningful information about microbial ecology. A further challenge is posed by the complexity and volume of data produced by HTS. A series of studies were conducted (**I**, **II**, **III**) to assess the importance of the sequencing platform, marker region and primer selection in capturing a wide range of AM fungal families. A final study (**IV**) led to the development of an easy-to-use pipeline incorporating many of the tools required by microbial ecologists and integrating the findings of papers **I–III**.

3.1. Intra- and interspecific variation of AM fungi in SSU, ITS and LSU marker regions

Paper **I** revealed that sequencing the SSU marker is a viable option along with ITS and LSU for AM fungal identification purposes. In eukaryotes, the SSU and LSU rRNA gene sequences are relatively conserved compared to the ITS region, due to their biological role of encoding ribosomal RNA (Abeyrathne and Nazar 2005). ITS regions, which are presumably under lower selective pressure, are expected to show higher variation. Within single AM fungal spores, almost every clone represented a new sequence variant in all the above mentioned three marker regions. Sequence variations were caused by substitutions and indels that generated up to 31bp variation in length for all regions. Average intraspecific sequence similarity values were 99%, 96% and 94% for amplicons in SSU, LSU and ITS, respectively. Nucleotide variation of all the SSU sequences showed that highest variation was apparent between 600–850 bases, which is located within the amplicon of widely used primers NS31-AML2 (Simon, Lalonde and Bruns 1992; Lee, Lee and Young 2008). The suitability of the central part of the SSU as a marker for AM fungal community surveys was further supported by its level of nucleotide variation, which is similar to that of the ITS region; its alignability across the entire phylum; its appropriate length for next-generation sequencing; and its ease of amplification in single-step PCR. Interspecific similarity could not be computed for ITS and LSU sequences, as these cannot be reliably aligned between families.

Sequences from spores were assigned to OTUs using the UCLUST 97% identity threshold OTU picking method in QIIME. Clustering showed that in most cases, more than one OTU was recovered for each morphospecies. Different clustering identity thresholds were used (95–99%) for three marker regions to

assess if automated OTU picking can be used for species identification (Figure 2). The number of generated OTUs was always higher for ITS and LSU than number of species, irrespective of the identity threshold.

These results show that automated OTU picking methods are sensitive to high levels of intraspecific sequence variation, which can cause overestimation of OTU richness in many applications. Similar OTU richness overestimation has been reported by others (Blaalid et al. 2013; Lindner et al. 2013; Sun et al. 2013). This can indicate that finding suitable general primers and applicable thresholds for automated OTU delimitation might be not possible (Ryberg 2015). BLAST-based sequence similarity showed that within spores, isolates and species the variation was lowest in SSU, intermediate in LSU and highest in ITS.
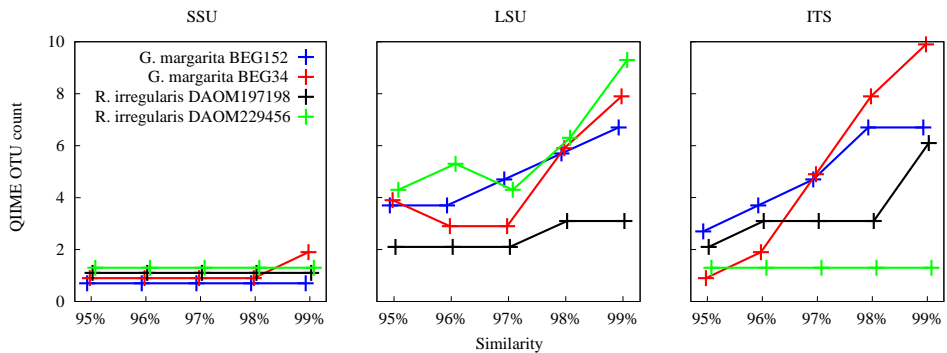

Figure 2. Number of Operational Taxonomic Units (OTUs) identified among SSU, ITS and LSU sequences sets derived from *Rhizophagus irregularis* isolates DAOM197198 and DAOM229456 and *Gigaspora margarita* isolates BEG34 and BEG152, at a range of sequence similarity thresholds. OTU picking was implemented in QIIME (Caporaso et al. 2010) using the default method (UCLUST). To improve readability of the graph the points for the different isolates are shifted upwards and rightwards. [Adapted by permission from Copyright Clearance Center: John Wiley and Sons Molecular Ecology "Sequence variation in nuclear ribosomal small subunit, internal transcribed spacer and large subunit regions of Rhizophagus irregularis and Gigaspora margarita is high and isolate-dependent", Thiéry et al. 2016, Copyright 2016]

3.2. Comparison of 454-sequencing and Illumina-sequencing using AM fungi

The SSU rRNA gene has variable and invariable regions, and has a relatively constant length (without many insertions or deletions) making it ideal for phylogenetic analysis and sequence alignment (**I**). When using the NS31-AML2 primer pair, the total length of the marker region is up to 560 bases, and most variable part is located after the forward primer (NS31) at 70 to 300 bases. In paper **II**, tagmentation based Illumina-sequencing, where obtained reads are located

randomly on the amplicon, overestimated richness. Using BLAST based indentification for the Illumina-sequencing without filtering amplicon location returned 216 VT compared to 52 VT obtained from the same samples when using 454-sequencing. By analyzing the VT BLAST hits it was apparent that the additional VT recorded in the unfiltered Illumina data represented reads located in the invariable part of the amplicon (300–520 nt). Further investigation with BLASTclust (Altschul et al. 1990) using 97% identity threshold indicated that 96.6% of these reads belonged to a single cluster, suggesting that BLAST-based identity assignment of similar score matches within the invariable region (300–520 nt) were largely random and correct hits could not be distinguished. We obtained relatively similar estimates of AM fungal diversity (richness) with Illumina MiSeq amplicon sequencing and 454-sequencing of the same samples, once reads were carefully quality filtered to match the variable region of the amplicon. This is notable given that Illumina MiSeq sequencing can provide two orders of magnitude greater sequencing depth than 454-sequencing but the sequencing depth is dependent on the success of the sequencing run and its configuration.

These results show that the increase in sequencing depth per sample with the Illumina approach did not further increase AM fungal diversity (richness) estimates per sample or total diversity estimates. We saw that Illumina-sequencing produced up to 30 000 reads per sample compared to the lowest count for 454-sequencing producing 300 reads per sample but observed AM fungal richness stayed similar (**II**). This is in strong contrast with the major improvement in the capture of AM fungal diversity that accompanied replacement of cloning-Sanger sequencing with 454-sequencing (Öpik et al. 2006; Öpik et al. 2009), and suggests that sufficient sequencing depth per sample (~300 sequences for AM fungi) has been reached. Appropriate quality filtering is crucial to avoid over-estimation of richness when using Illumina MiSeq sequencing and BLAST-based bioinformatics for read identification (Figure 3). While using BLAST-based bioinformatics, marker properties need to be considered and only sufficiently variable marker regions should be used. Rarefaction analysis suggested that the number of AM fungal reads per sample for 454-sequencing and Illumina-sequencing was generally sufficient to produce asymptotic estimates of VT richness per sample. The 10 most abundant shared VT for both sequencing methods represented 74.8% and 88% of reads for amplicon filtered 454 and amplicon filtered Illumina sequence sets, respectively. There was no correlation between dry root weight and sequence depth or recorded AM fungal richness.
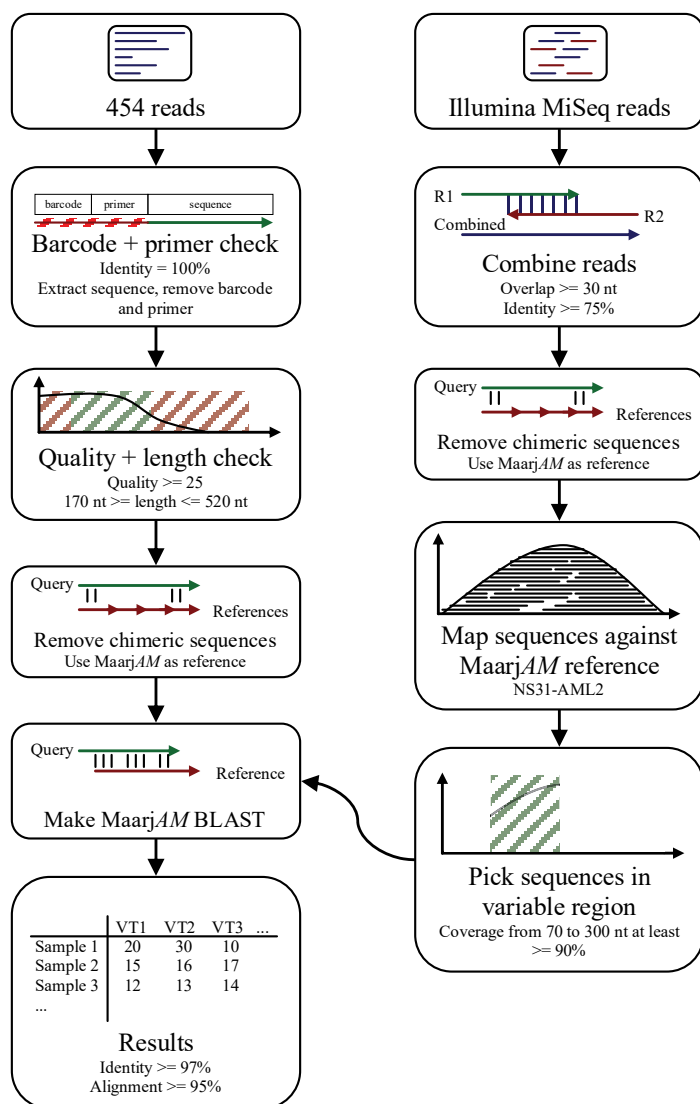
Figure 3. Pipeline workflow of 454-sequencing and Illumina-sequencing used in paper **II**.

Metagenomic (PCR-free) Illumina-sequencing of plant root DNA in paper **II** returned only 309 AM fungal hits and 16 VT from 21 640 048 sequences, indicating that a very low proportion of AM fungal SSU rRNA gene sequences is detectable in plant root metagenomes. From metagenomic data, 41 732 potential genes were found, and protein BLAST returned 2728 protein sequences, of which 88% were of plant origin and 11% were fungal proteins. This can be an indication that only a small number of species are represented in the protein sequence database and the amount of AM fungal DNA in root samples is low compared to that of the host plant and other organisms living in the plant. This suggests that amendments

to the metagenomic sequencing approach are required for a PCR-free description of AM fungal assemblages to become feasible.

Metagenomic sequencing would help to solve biases caused by PCR and would avoid generation of chimeric sequences originating in the PCR step. However, at the current stage it is not a viable approach for community analyses, as the amount of fungal DNA compared to host plant and other organisms in the soil is minute and only a small fraction of the fungal sequences actually present in samples can be recovered (**II**). Whole genome sequencing projects have improved the species coverage in the genome databases and available programs with information from databases has extended the knowledge of retrieved genes, which can help to identify and understand underlying functions and expressions of different organism groups and their interactions. Still there is lack of information as the sequencing has completed for only a fraction of species and the genomes often contain a lot of gaps and non-sequenced regions (Nakamura et al. 2011; Santamaria et al. 2012; del Campo et al. 2014; Boon et al. 2015; Morin et al. 2019). Although sequencing cost has decreased, the whole genome sequencing is still expensive and frequently does not satisfy research needs for microbial community ecology. Current identification of organisms from metagenomics data relies on gene prediction found from these genome sequences, which might still miss crucial information or can be wrongly annotated.

3.3. AM fungal richness using ITS and SSU amplification

Paper **III** tested and validated the performance of ITS2 region amplified with general fungal primers in comparison with parallel sample analysis using AM fungal specific SSU rRNA gene primers. All of the samples contained AM fungal sequences but yielded different proportions of particular AM fungi. Both markers detected similar community responses to experimental treatments. This may indicate that both target regions amplified the abundant Glomeraceae family in a similar manner as indicated by significant positive correlation in SSU and ITS Glomeraceae sequence abundance across samples in both studies ($R_{Greenhouse}$ = 0.73, $P<0.001$; R_{Field} = 0.76, P<0.001). More than 90% of quality filtered reads of the SSU dataset used in the greenhouse and in the field experiment were AM fungal reads, matched with the Maarj*AM* database. In contrast, the proportion of AM fungal sequences was substantially lower in the two ITS datasets obtained with general fungal primers. The ITS sequence data from the greenhouse experiment samples had only 11% of total sequences matched with Maarj*AM* ITS sequences compared to 43% from the field survey. The difference between the two datasets shows that the proportion of AM fungi amplified by general fungal primers can vary among samples, studies and host plant species. Even though AM fungal sequences often constitute a relatively small proportion of the total fungal sequence data in samples, general fungal ITS primers and HTS can still characterize AM fungal diversity and the most abundant taxa. ITS primers aiming to amplify all fungi have been constantly improved to capture a wider range of

clades and species within kingdom Fungi, including improved detection of Glomeromycotina sequences. Both markers reported highly significant (P=0.001 in all cases) responses of AM fungal community composition to soil moisture (R_{SSU}=0.36; R_{ITS}=0.20) and season (R_{SSU}=0.20; R_{ITS}=0.31).

Paper **III** showed that the general fungal ITS amplicon returned fewer AM fungal sequences compared to the AM fungal specific SSU amplicon but, as expected, yielded more other fungi. For studying AM fungal diversity patterns and focusing on abundant taxa, the ITS with improved general fungal primers can be an appropriate candidate in place of the SSU, because they capture other fungi in the soil or roots. However, for an all-inclusive study multiple primers and regions may need to be used to overcome biases towards specific organism groups (Lentendu et al. 2014; Heise et al. 2015).

All current approaches used to characterize AM fungal communities have trade-offs that should be considered when selecting markers. For example, AM fungal-specific SSU primers may be able to amplify more families and provide a more detailed view of AM fungal communities than general fungal ITS primers. However, if the main goal of a study is to assess treatment responses of mainly dominant taxa, then our results indicate that general fungal ITS primers could be used, as long as a sufficient number of Glomeromycotinan sequences are obtained in all samples to detect adequate completeness of the existing diversity. Using general fungal ITS primers is advantageous as they can reveal community-level responses and interactions across the kingdom Fungi using the same effort and resources as targeting AM fungi alone with SSU primers.

3.4. Summary of the pipeline

gDAT (Figure 4), the pipeline developed in this study, has been already used to analyse multiple environmental HTS datasets to assess AM fungal communities (Davison et al. 2015; García de León et al. 2018). The HTS data for papers **II** and **III** were analysed using the earlier versions of gDAT, showing that development has been ongoing to keep pace with current sequencing technologies and the needs of users. The pipeline consists of a wide package of open-source bioinformatics software and custom python scripts to process data into the correct formats and summarize each software output (Table 2). This conversion and immediate experience allow users to quickly analyse HTS data with little turnover time. Each downstream analysis step in the pipeline is offered with multiple options to allow optimal data processing and analysis for the dataset used. gDAT can process raw sequencing data from multiple sequencing platforms including Sanger, 454, IonTorrent, Illumina and PacBio sequencing. The pipeline is made available for multiple operating systems allowing the end-user to get quickly started with data analyses. The pipeline is packaged with the Maarj*AM* database as a reference dataset that can be immediately used to analyse environmental AM fungal datasets. gDAT thus facilitates comparability between studies and more robust metastudies, as datasets can be cleaned in a similar fashion and use the same molecular

taxonomic units from Maarj*AM* database (VT). Experiments in paper **IV** showed that analyses can be conducted within a day on a commodity hardware with at least 50 GB free disk space. CPU and memory consumption monitoring revealed that clustering, chimera checking, and identification can be sped up by having multiple cores, but the most memory demanding process was sequence clustering. Also, general eukaryotic primers were able to capture a portion of the most abundant AM fungal sequences from the samples.
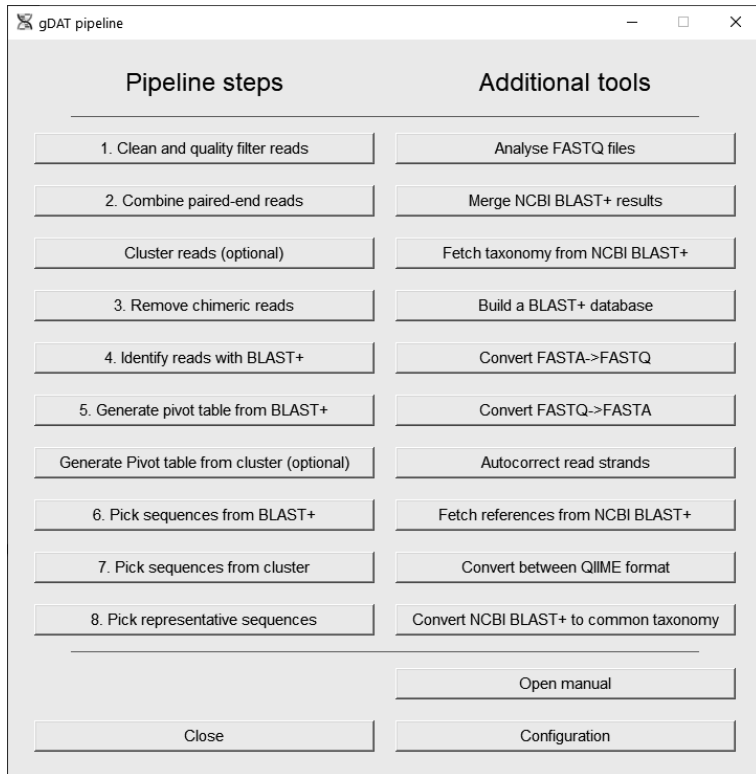

Figure 4. Screenshot from gDAT pipeline home screen showing various options and numbered analysis steps for easy tracking. Each step can be executed at any time and the pipeline can be stopped and restarted from the position where it was halted.

3.4.1. BLAST-based sequence identification

The probability of a correct identity assignment may also be increased by BLAST hit filtering on the basis of alignment length to include only hits with approximately full-length alignment and remove chimeric reads (**II**). BLAST will work optimally if the query and reference sequences are trimmed to the same position on the amplicon and to the same length. Still the robustness of the BLAST identification process was shown in (**II**) where reference sequences were consistently identified before and after artificial trimming (**II**). BLAST heavily relies on

flanked conservative regions, as mismatches in these regions can result in shorter alignments to maximize BLAST score, making alignment-based criteria problematic as we might misplace hits and record lower species richness. This can be problematic when using the ITS region. The pipeline offers one option to overcome the problem by allowing the inclusion of the primer sequences, which otherwise are removed, when cleaning the sequences. This allows the conserved portion of the sequence to be retained, allowing better alignment with the BLAST algorithm.

3.4.2. Available databases

For each marker region different databases can be used to identify sequences. Suitable reference databases should be chimera free and correctly annotated to minimise misidentifications in studies. When using conserved regions that are easy to align between families, phylogenetic analysis can be used to determine the correct identification of reference sequences (Öpik et al. 2010). This is mostly manual work and is feasible with smaller databases or database partitions. There is an option to use a clustering-based approach and only align centroids (cluster center points) for phylogenetic analysis to decrease the size of the tree. Also, large trees tend to be noisy, as small variations and multiple indels are challenging for alignment algorithms, making it difficult to distinguish generated clades correctly. For SSU we used the Maarj*AM* database, which contains AM fungal sequences covering SSU, ITS and LSU regions. The Maarj*AM* database also defines SSU-based VT, which are potential species delimited by phylogenetic analyses (**II**, **III**, **IV**). For ITS we have used both UNITE and Maarj*AM* databases to identify sequences (**III**, **IV**). Other similar curated specific databases include SILVA (containing SSU and LSU sequences for Bacteria, Archaea and Eukarya; Quast et al. 2012), PHYMYCO-DB (containing SSU and EF1-α sequences for Fungi; Mahé et al. 2012) and ITSoneDB (containing ITS1 sequences for Eukarya; Santamaria et al. 2018).

However, it should be noted that manually curated databases are usually not as inclusive as general databases, despite the curators' best efforts. Incompleteness of reference sequence databases and unresolved taxonomic relationships complicates taxonomic placement of fungal sequences. Sequences that are not matched against a target organism in the reference database (called no-hits) can be further subjected to a search against the INSDC database (Cochrane, Karsch-Mizrachi and Takagi 2016). The INSDC database provides a valuable free asset for scientists, but comes with a certain degree of mistrust, as it contains misidentified, chimeric, reverse complement and low-quality sequences as well as insufficient metadata (Tedersoo et al. 2011; Xu 2016; Nilsson et al. 2017). Using hits from the INSDC database can provide a rough estimate of what is present in a sample, but does not provide very reliable ecological information, as it is likely to contain a lot of noise, while some species can have very small interspecific variation depending on the marker region (Cabelin and Alejandro 2016). To overcome

these issues, the developed pipeline supports the usage of multiple best BLAST hits to construct a common taxonomy for each no-hit, with a certain degree of trust based on the taxonomic rank revealed (**IV**). The approach of using a common taxonomic rank for each sequence cluster produces reliable estimates of taxonomic identity for clusters but has the caveat of producing a nested taxonomy that contains some very diverse taxa and others with very fine resolution. The exact lists of plant and soil animal taxa that is returned should be interpreted in this light and taken only as a rough indication of which groups of taxa were present. These no-hits have provided valuable information while updating Maarj*AM* database with missing records. Hits marked as potential Glomeromycotina might represent putative new VT that can be added to the Maarj*AM* database but need to be verified in each case. Phylogenetic analysis is used to confirm new clades or addition of the new reference sequences to an existing clade. Since no-hits provide information about other, non-target organisms the selected primers have detected, this may indicate directions for troubleshooting in the case of problems related to low sequence counts or low hit counts while using a targeted reference database.

3.4.3. Commodity hardware

To speed up the performance of the pipeline, gDAT incorporates an option to precluster cleaned sequences into smaller sets to reduce computational time and complexity, which is handy for commodity hardware. As clustering and chimera checking programs often work only with sequences being in the same strand, gDAT also allows to reverse complement incorrectly orientated reads (**IV**). However, BLAST allows to have both orientations and the outcome is not affected. gDAT produces lookup files to map chimera checking results and BLAST identifications back to individual sequences. A caveat of the clustering is that proper identity threshold to separate clusters has to be used, otherwise in the case of lower threshold some closely related taxa can be clustered together reducing taxon richness, or when using higher threshold value, the richness can be overestimated. Another issue is with heuristic based clustering methods (greedy) that depend on the order of the sequences and can produce different outcomes if the input files have been slightly changed (discussed by Kellom and Raymond 2017). This is not an issue with single-linkage clustering (hierarchical), but as the sequencing depth has increased tremendously, this approach is not feasible as it is highly resource and time-consuming to process the reads. To overcome the problem, increased identity threshold compared to BLAST identification threshold can be used for heuristic based clustering to allow separation of closely related taxa and correct BLAST identification of the reads while mapping back from cluster centroids to individual reads. For ITS and SSU, where sequence identification to species level is roughly correct using 97% identity in the case of AM fungi, a clustering level of at least 98% identity or higher has to be used.

4. CONCLUSIONS

The culmination of this thesis was the production of an easy-to-use bioinformatic pipeline to analyse biodiversity data. AM fungal samples from spores, roots and soil were also identified using different marker regions and sequencing platforms to improve our understanding of how unculturable organisms can be studied. The pipeline gDAT has already proven to be a valuable tool for ecologists. The tools within gDAT also reduce input errors and potentially unify analyses across studies by suggesting recommended thresholds for each cleaning and analysing step. Unified analyses allow data from previous studies to be used as metadata as the common pipeline and database result in comparable outcomes. Building and improving the pipeline has required a constant effort to modify analysis parameters and add new implementation to handle HTS data from new sequencing platforms. Modifying the parameters allows broader or narrower targeting of organism groups and marker regions in analyses as needed.

The following main conclusions can be inferred from the thesis:
- The SSU rRNA gene region provides sufficient resolution to recognise AM fungal species level taxa (**I**).
- Different high-throughput sequencing (HTS) platforms detect similar community patterns but exhibit some technical differences: they detect the same abundant taxa but show differences in terms of rare taxon detection (**II**).
- Careful bioinformatic filtering has to be applied to HTS data to obtain biologically meaningful diversity estimates (**II**).
- Further increasing sequencing depth per sample of different HTS approaches does not detect higher species richness in the case of AM fungi; though this was the case when switching from Sanger-sequencing to 454-sequencing, where a 100-fold increase in sequencing depth per sample was achieved (**II**).
- Available primer combinations targeting different marker regions are able to capture most abundant AM fungal families yet yield different community composition. However, they detect similar community dynamics in response to changing soil moisture or across seasons (**III**).
- The software gDAT allows efficient and rapid analysis of AM fungal soil and root amplicon data using various marker regions and sequencing technologies (**IV**).

SUMMARY

A major goal of ecology is to understand diversity patterns and underlying processes. Arbuscular mycorrhizal (AM) fungi are one of the key microbial groups in grassland and deciduous forest ecosystems, providing nutrients to host plants, protecting host plants from pathogens and increasing drought resistance. Studying these systems helps us to understand the mechanisms and symbioses between these two organism groups. As AM fungi are often unculturable (hard to maintain in the lab environment) and are obligate symbionts, studying these microorganisms is extremely challenging. AM fungi produce spores and hyphae that can be gathered from roots or soil and can be studied microscopically. However, the morphological characters of AM fungi often vary little even between families and need taxonomic expertise to identify them correctly. Advances in molecular methods (DNA sequencing) have greatly helped to identify these species and improved studies of natural ecosystems. The costs of sequencing have dropped to the extent that global-scale studies are feasible. As these microorganisms contribute a very low volume of DNA to the soil metagenome, the most cost-effective way to target them is to amplify specific marker regions. In this thesis I considered three major marker regions used in AM fungal sequencing studies and found that small subunit (SSU) rRNA gene marker is optimal as it provides sufficient variation, has a constant length and allows a wide variety of organisms present in roots and soil to be captured.

For identification purposes, we also need a reference database, where reference sequences corresponding to the employed marker region are provided for a variety of species. If we are dealing with a new marker region or new organisms, such databases need to be created and constantly curated. A considerable amount of work has been put into developing the Maarj*AM* database, which contains AM fungal reference sequences from SSU, ITS and LSU marker regions. As no database is ideal and does not cover all species, additional databases need to be used or unidentified sequences need to be identified either by distinguishing new species with the help of a microscope or by mapping unidentified sequences onto a phylogenetic tree of known species. Large data volumes and increasing sequence output per study has generated a new bioinformatic problem where in-house expertise is needed to complete the data analysis. Often it is impractical for workgroups to employ bioinformatics specialists and they therefore cannot conduct such analyses at the optimal level for specific research projects. My thesis focuses on producing software (a pipeline called gDAT) to help ecologists conduct bioinformatic data analysis with ease. Providing a graphical interface and an optimal configuration with predefined values for each field allows scientists to focus on drawing conclusions from their data rather than focusing on learning the new software and bioinformatics tools. The pipeline is provided with AM fungal reference SSU rRNA gene sequences from the Maarj*AM* database. The pipeline is easy to use also with other reference databases, other markers and hence for other groups of organisms.

The main results of the thesis are the following: 1) the SSU marker region provides sufficient resolution to distinguish different AM fungal species, 2) newer sequencing platforms like Illumina provide similar AM fungal species richness estimates compared with older platforms like 454, 3) using a different marker region (ITS) and broader primer specificity produces similar AM fungal community patterns and responses to seasons and soil moisture treatments compared to analysis based on the AM fungal specific primers for SSU rRNA gene region and 4) high-throughput sequencing of amplicon data provides a means to analyse AM fungal sequences from soil and plant roots, implemented in easy, flexible and user-friendly manner in bioinformatical pipeline gDAT.

REFERENCES

Abarenkov K, Henrik Nilsson R, Larsson KH, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R. 2010. The UNITE database for molecular identification of fungi–recent updates and future perspectives. *New Phytologist*, 186(2):281–285.

Abeyrathne PD, Nazar RN. 2005. Parallels in rRNA processing: conserved features in the processing of the internal transcribed spacer 1 in the pre-rRNA from Schizosaccharomyces pombe. *Biochemistry*, 44(51):16977–16987.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Anslan S, Bahram M, Hiiesalu I, Tedersoo L. 2017. PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Molecular Ecology Resources*, 17(6):e234–240.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research*, 41:D36-D42.

Blaalid R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kauserud H. 2013. ITS 1 versus ITS 2 as DNA metabarcodes for fungi. *Molecular Ecology Resources*, 13(2):218–224.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics 30*:2114–2120.

Boon E, Halary S, Bapteste E, Hijri M. 2015. Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm. *Genome Biology and Evolution*, 7(2):505–521.

Burke CM, Darling AE. 2016. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ*, 4:e2492.

Cabelin VL, Alejandro GJ. 2016. Efficiency of matK, rbcL, trnH-psbA, and trnL-F (cpDNA) to molecularly authenticate Philippine ethnomedicinal Apocynaceae through DNA barcoding. *Pharmacognosy Magazine*, 12(Suppl 3):S384.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38:1767–1771.

Cochrane G, Karsch-Mizrachi I, Takagi T, Sequence Database Collaboration IN. 2016. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 44(D1):D48-D50.

Corradi N, Brachmann A. 2017. Fungal mating in the most widespread plant symbionts? *Trends in Plant Science*, 22(2):175–183.

Cui XC, Hu JL, Wang JH, Yang JS, Lin XG. 2016. Reclamation negatively influences arbuscular mycorrhizal fungal community structure and diversity in coastal saline-alkaline land in Eastern China as revealed by Illumina-sequencing. *Applied Soil Ecology*, 98:140–149.

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, Johnson NC, Kane A, Koorem K, Kochar M, Ndiaye C, Pärtel M, Reier Ü, Saks Ü, Singh R, Vasar M, Zobel M. 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science*, 349:970–973.

Davison J, Öpik M, Zobel M, Vasar M, Metsis M, Moora M. 2012. Communities of arbuscular mycorrhizal fungi detected in forest soil are spatially heterogeneous but do not vary throughout the growing season. *PLoS One*, 7:e41938.

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. 2014. The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, 29:252–259.

Dumbrell AJ, Ashton PD, Aziz N, Feng G, Nelson M, Dytham C, Fitter AH, Helgason T. 2011. Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytologist*, 190(3):794–804.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27:2194–2200.

García de León D, Davison J, Moora M, Öpik M, Feng H, Hiiesalu I, Jairus T, Koorem K, Liu Y, Phosri C, Sepp SK. 2018. Anthropogenic disturbance equalizes diversity levels in arbuscular mycorrhizal fungal communities. *Global Change Biology*, 24(6):2649–59.

Gollotte A, van Tuinen D, Atkinson D. 2004. Diversity of arbuscular mycorrhizal fungi colonising roots of the grass species Agrostis capillaris and Lolium perenne in a field experiment. *Mycorrhiza*, 14(2):111–117.

Gordon ND, McMahon TA, Finlayson BL, Gippel CJ, Nathan RJ. 2004. Stream hydrology: an introduction for ecologists. John Wiley and Sons.

Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PloS One*, 9(2):e87624.

Hart MM, Aleklett K, Chagnon PL, Egan C, Ghignone S, Helgason T, Lekberg Y, Öpik M, Pickles BJ, Waller L. 2015. Navigating the labyrinth: A guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytologist*, 207:235–247.

Heise W, Babik W, Kubisz D, Kajtoch Ł. 2015. A three-marker DNA barcoding approach for ecological studies of xerothermic plants and herbivorous insects from central Europe. *Botanical Journal of the Linnean Society*, 177(4):576–592.

Ihrmark K, Bödeker I, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE, Lindahl BD. 2012. New primers to amplify the fungal ITS2 region–evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*, 82(3):666–677.

Johansen RB, Johnston P, Mieczkowski P, Perry GLW, Robeson MS, Burns BR, Vilgalys R. 2016. A native and an invasive dune grass share similar, patchily distributed, root-associated fungal communities. *Fungal Ecology*, 23:141–155.

Kellom M, Raymond J. 2017. Using cluster edge counting to aggregate iterations of centroid-linkage clustering results and avoid large distance matrices. *Journal of Biological Methods*, 4(1):e68.

Kivlin SN, Hawkes CV, Treseder KK. 2011. Global diversity and distribution of arbuscular mycorrhizal fungi. *Soil Biology and Biochemistry*, 43(11):2294–2303.

Kohout P, Sudová R, Janoušková M, Čtvrtlíková M, Hejda M, Pánková H, Slavíková R, Štajerová K, Vosátka M, Sýkorová Z. 2014. Comparison of commonly used primer sets for evaluating arbuscular mycorrhizal fungal communities: Is there a universal solution? *Soil Biology and Biochemistry*, 68:482–493.

Lee HK, Lee CK, Tang JW, Loh TP, Koay ES. 2016. Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. *Scientific Reports*, 6:33318.

Lee J, Lee S, Young JPW. 2008. Improved PCR primers form the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiology Ecology*, 65:339–349.

Lekberg Y, Schnoor T, Kjøller R, Gibbons SM, Hansen LH, Al-Soud WA, Sørensen SJ, Rosendahl S. 2012. 454-sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. *Journal of Ecology*, 100(1):151–160.

Lentendu G, Wubet T, Chatzinotas A, Wilhelm C, Buscot F, Schlegel M. 2014. Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. *Molecular Ecology*, 23(13):3341–3355.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452.

Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, Pang E, Cao H, Cha H, Lin T, Zhou Q, Shang Y, Li Y, Sharma T, van Velzen R, de Ruijter N, Aanen DK, Win J, Kamoun S, Bisseling T, Geurts R, Huang S. 2014. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PloS Genetics*, 10:e1004078.

Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjoller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J, Kauserud H. 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist*, 199:288–299.

Lindner DL, Carlsen T, Henrik Nilsson R, Davey M, Schumacher T, Kauserud H. 2013. Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, 3(6):1751–1764.

Lumini E, Orgiazzi A, Borriello R, Bonfante P, Bianciotto V. 2010. Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. *Environmental Microbiology*, 12(8):2165–2179.

Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27:2957–2963.

Mahé S, Duhamel M, Le Calvez T, Guillot L, Sarbu L, Bretaudeau A, Collin O, Dufresne A, Kiers ET, Vandenkoornhuyse P. 2012. PHYMYCO-DB: a curated database for analyses of fungal diversity and evolution. *PLoS One*, 7(9):e43117.

Meyer CP, Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, 3(12):e422.

Moora M, Berger S, Davison J, Öpik M, Bommarco R, Bruelheide H, Kühn I, Kunin WE, Metsis M, Rortais A, Vanatoa A. 2011. Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454-sequencing. *Journal of Biogeography*, 38(7):1305–1317.

Morin E, Miyauchi S, San Clemente H, Chen EC, Pelin A, de la Providencia I, Ndikumana S, Beaudet D, Hainaut M, Drula E, Kuo A, Tang N, Roy S, Viala J, Henrissat B, Grigoriev IV, Corradi N, Roux C, Martin FM. 2019. Comparative genomics of Rhizophagus irregularis, R. cerebriforme, R. diaphanus and Gigaspora rosea highlights specific genetic features in Glomeromycotina. *New Phytologist*, 222(3):1584–1598.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39:e90.

Nilsson RH, Larsson KH, Taylor AF, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, 47(D1):D259-D264.

Nilsson RH, Sánchez-García M, Ryberg MK, Abarenkov K, Wurzbacher C, Kristiansson E. 2017. Read quality-based trimming of the distal ends of public fungal DNA sequences is nowhere near satisfactory. *MycoKeys*, 26:13.

Noreika N, Helm A, Öpik M, Jairus T, Vasar M, Reier Ü, Kook E, Riibak K, Kasari L, Tullus H, Tullus T, Lutter R, Oja E, Saag A, Randlane T, Pärtel M. 2019. Forest biomass, soil and biodiversity relationships originate from biogeographic affinity and direct ecological effects. *Oikos*, 128(11):1653–1665.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin RP, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2012. vegan: community ecology package, version 2.0-4. URL http://CRAN.R-project.org/package=vegan.

Oksanen J, Blanchet GF, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2016. vegan: Community Ecology Package. R package version 2.3-5. https://CRAN.R-project.org/package=vegan.

Öpik M, Davison J, Moora M, Zobel M. 2014. DNA-based detection and identification of Glomeromycota: The virtual taxonomy of environmental sequences. *Botany*, 92: 135–147.

Öpik M, Davison J. 2016. Uniting species- and community-oriented approaches to understand arbuscular mycorrhizal fungal diversity. *Fungal Ecology*, 24B:106–113.

Öpik M, Metsis M, J T, Zobel M, Moora M. 2009. Large-scale parallel 454-sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist*, 184:424–437.

Öpik M, Moora M, Liira J, Zobel M. 2006. Composition of root-colonizing arbuscular mycorrhizal fungal communities in different ecosystems around the globe. *Journal of Ecology*, 94:778–790.

Öpik M, Moora M, Zobel M, Saks Ü, Wheatley R, Wright F, Daniell T. 2008. High diversity of arbuscular mycorrhizal fungi in a boreal herb-rich coniferous forest. *New Phytologist*, 179:867–876.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database Maarj*AM* reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, 188:223–241.

Öpik M, Zobel M, Cantero JJ, Davison J, Facelli JM, Hiiesalu I, Jairus T, Kalwij JM, Koorem K, Leal ME, Liira J, Metsis M, Neshataeva V, Paal J, Phosri C, Põlme S, Reier Ü, Saks Ü, Schimann H, Thiéry O, Vasar M, Moora M. 2013. Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza*, 23:411–430.

Orchard S, Hilton S, Bending GD, Dickie IA, Standish RJ, Gleeson DB, Jeffery RP, Powell JR, Walker C, Bass D, Monk J, Simonin A, Ryan MH. 2017. Fine endophytes (Glomus tenue) are related to Mucoromycotina, not Glomeromycota. *New Phytologist*, 213:481–486.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Resear*ch, 41(D1):D590–D596.

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

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12:85–94.

Ryberg M. 2015. Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. *Molecular Ecology*, 24(23):5770–5777.

Saks Ü, Davison J, Öpik M, Vasar M, Moora M, Zobel M. 2014. Root-colonizing and soil-borne communities of arbuscular mycorrhizal fungi in a temperate forest understory. *Botany*, 92:277–285.

Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G, Pesole G. 2012. Reference databases for taxonomic assignment in metagenomics. *Briefings in Bioinformatics*, 13:682–695.

Santamaria M, Fosso B, Licciulli F, Balech B, Larini I, Grillo G, De Caro G, Liuni S, Pesole G. 2018. ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences. *Nucleic Acids Research*, 46(D1): D127–D132.

Sepp SK, Davison J, Moora M, Neuenkamp L, Oja J, Öpik M, Roslin T, Vasar M, Zobel M. 2020. Complex changes in the functional structure of the above- and belowground biotic community with woody encroachment in a grassland. Manuscript submitted for publication.

Sepp SK, Jairus T, Vasar M, Zobel M, Öpik M. 2018. Effects of land use on arbuscular mycorrhizal fungal communities in Estonia. *Mycorrhiza*, 28(3):259–268.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21:1794–1805.

Simon L, Lalonde M, Bruns TD. 1992. Specific amplification of 18S fungal ribosomal genes from vesicular-arbuscular endomycorrhizal fungi colonizing roots. *Applied and Environmental Microbiology*, 58:291–295.

Smith SE, Read DJ. 2008. Mycorrhizal symbiosis. Academic Press. Amsterdam.

Spatafora JW, Chang Y, Benny GL, Lazarus KL, Smith ME, Berbee ML, Bonito G, Corradi N, Grigoriev I, Gryganskyi A, James TY, O'Donnell K, Roberson RW, Taylor TN, Uehling J, Vilgalys R, White MM, Stajich JE. 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia*, 108:1028–1046.

Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Applied and Environmental Microbiology*, 79:5962–5969.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8):2045–2050.

Tamura K, Dudley J, Nei M, Kumar S. 2007. Mega4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24:1596–1599.

Taylor JD, Helgason T, Öpik M. 2017. Molecular community ecology of arbuscular mycorrhizal fungi. In: Dighton J, White JF, eds. The Fungal Community: Its Organization and Role in the Ecosystem, 4th edn. *CRC Press*, pp 1–26.

Tedersoo L, Abarenkov K, Nilsson RH, Schüssler A, Grelet GA, Kohout P, Oja J, Bonito GM, Veldre V, Jairus T, Ryberg M. 2011. Tidying up international nucleotide sequence databases: ecological, geographical and sequence quality annotation of ITS sequences of mycorrhizal fungi. *PLoS One*, 6(9):e24940.

Tedersoo L, Bahram M, Zobel M. 2020. How mycorrhizal associations drive plant population and community biology. *Science*, 367(6480):eaba1223.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G and Kõljalg U. 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytolo*gist, 188:291–301.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research*, 18:1979–1990.

Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P, Duensing N, Frei dit Frey N, Gianinazzi-Pearson V, Gilbert LB, Handa Y, Herr JR, Hijri M, Koul R, Kawaguchi M, Krajinski F, Lammers PJ, Masclaux FG, Murat C, Morin E, Ndikumana S, Pagni M, Petitpierre D, Requena N, Rosikiewicz P, Riley R, Saito K, San Clemente H, Shapiro H, van Tuinen D, Becard G, Bonfante P, Paszkowski U, Shachar-Hill Y, Tuskan GA, Young JPW, Sanders IR, Henrissat B, Rensing SA, Grigoriev IV, Corradi N, Roux C, Martin F. 2013. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences*, 110:20117–20122.

Treseder KK. 2016. Model behavior of arbuscular mycorrhizal fungi: Predicting soil carbon dynamics under climate change. *Botany*, 94:417–423.

van der Heijden MGA, Bardgett RD, van Straalen NM. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology Letters*, 11:296–310.

van der Heijden MGA, de Bruin S, Luckerhoff L, van Logtestijn RSP, Schlaeppi K. 2015. A widespread plant-fungal-bacterial symbiosis promotes plant biodiversity, plant nutrition and seedling recruitment. *The ISME Journal*, 10:389–399.

van Tuinen D, Jacquot E, Zhao B, Gollotte A, Gianinazzi-Pearson V. 1998. Characterization of root colonization profiles by a microcosm community of arbuscular mycorrhizal fungi using 25S rDNA-targeted nested PCR. *Molecular Ecology*, 7(7):879–887.

Varela-Cervero S, Vasar M, Davison J, Barea JM, Öpik M, Azcón-Aguilar C. 2015. The composition of arbuscular mycorrhizal fungal communities differs among the roots, spores and extraradical mycelia associated with five Mediterranean plant species. *Environmental Microbiology*, 17(8):2882–2895.

Wang C, White PJ, Li CJ. 2016. Colonization and community structure of arbuscular mycorrhizal fungi in maize roots at different depths in the soil profile respond differently to phosphorus inputs on a long-term experimental site. *Mycorrhiza*, 27:369–381.

Wang Y, Qian PY. 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PloS One*, 4(10):e7401.

White TJ, Bruns T, Lee SJWT, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. PCR Protocols: A Guide to Methods and Applications. *San Diego: Academic Press, Inc*., 315–322.

Xu TL, Veresoglou SD, Chen YL, Rillig MC, Xiang D, Ondrej D, Hao ZP, Liu L, Deng Y, Hu YJ, Chen WP, Wang JT, He JZ, Chen BD. 2016. Plant community, geographic distance and abiotic factors play different roles in predicting AMF biogeography at the regional scale in northern China. *Environmental Microbiology Reports*, 8:1048–1057.

SUMMARY IN ESTONIAN

Bioinformaatilise programmi gDAT arendus arbuskulaarset mükoriisat moodustavate seente määramiseks erinevate pärilikkusaine markerpiirkondade põhjal

Ökoloogia uurib organismide omavahelisi interaktsioone, organismide leviku-mustreid ja nende seoseid keskkonnaga. Mullas elab palju mikroorganisme, kes täidavad olulisi rolle ökosüsteemide toimimises, näiteks aineringete mõjutajatena, keskkonna kujundajatena ning organismide kasvu ja omavaheliste seoste suuna-jatena. Üheks oluliseks mikroorganismide rühmaks mullas on arbuskulaarset mükoriisat (AM) moodustavad seened (krohmseened). AM on seenjuure (taime-juure ja seente kooselu) üks vormidest, mida moodustavad krohmseened enamuse roht- ja puittaimedega, sealhulgas paljude kultuurtaimedega. Mükoriisses kooselus saab peremeestaim seene vahendusel enda kasvuks vajalikke mineraalaineid ja vett, seen omakorda saab taimelt fotosünteesi käigus fikseeritud süsivesikuid. Lisaks parandavad AM seened taimede toimetulekut stressitingimustega, näiteks veepuuduse ja haigustekitajatega. AM seentega elab sümbioosis hinnanguliselt enam kui 80% soontaimedest ning seda esineb pea kõigis maismaa ökosüsteemi-des üle maailma.

Enne molekulaarsete meetodite kasutusele võtmist sai mikroorganisme teha kindlaks ja määrata morfoloogiliste ja füsioloogiliste tunnuste põhjal. See on lihtsamini teostatav AM seeneliikidel, mida on võimalik kultuuri eraldada ning laboritingimustes kasvatada. AM seeni on laboritingimustes keerukas kasvatada, sest nad on obligatoorsed sümbiondid taimedega; neil on eluspüsimiseks vaja peremeestaime. Laboris saab neid kasvatada vaid kultuuris koos taimedega, mis on keerukas ning kõik seeneliigid sellistes tingimustes hästi ei kasva. AM seente keha koosneb peenikesest seeneniidistikust mullas ning taimejuurtes ning eostest, vesiikulitest ja arbuskulitest taimejuurtes. Need struktuurid erinevad liikide vahel vähesel määral ja nii ongi valdavalt eosed, mille tunnuste alusel klassikaliselt AM seeneliike omavahel eristatakse.

Molekulaarsete meetodite, eriti pärilikkusaine (DNA) järjestamine, ja tehno-loogiate areng on teinud võimalikuks uurida AM seeni ja teisi mikroorganisme nii mullas kui ka taimejuurtes, ning looduslikes ja tehislikes ökosüsteemides (näiteks põllul, õhus ja tolmus). DNA järjestamise metoodika aitab uurida AM seente elu ja toimimist keskkonnas. Molekulaarsete meetodite edukaks kasuta-miseks on vaja välja töötada vahendid, mis sobivad konkreetse organismirühma, näiteks AM seente tõhusaks määramiseks. AM seente uurimisel kasutatakse liikide määramiseks peamiselt genoomi osi, mis kodeerivad ribosomaalse RNA ühikuid, näiteks ribosomaalse RNA väikest alaühikut (SSU), suurt alaühikut (LSU) ning sama genoomipiirkonna mittekodeerivat osa ITS. Vastavate markerpiirkondade nukleotiidsete järjestuste omavahelise sarnasuse võrdlemiseks ning liigilise kuuluvuse määramiseks saab kasutada erinevaid järjestuste võrdluse programme ja algoritme (näiteks BLAST, Smith-Waterman, Needleman-Wunsch, usearch).

Erinevate algoritmide kasutamine järjestuste määramisel eeldab referentsandmebaasi olemasolu, mis sisaldaks piisavalt järjestusi erinevatest taksonoomilistest üksustest (liigid, perekonnad jne); need peavad olema kontrollitud ja ekspertide poolt määratud ning annoteeritud. Oma doktoritöös uurisin, kuidas neid raskesti uuritavaid seeni saab määrata DNA järjestuste alusel ning selle käigus arendasin bioinformaatilise programmi, mis aitab teostada pärilikkusaine põhiseid AM seente elurikkuse uuringuid.

Krohmseente elurikkuse andmebaas Maarj*AM* on välja töötatud Tartu Ülikoolis. See sisaldab AM seente DNA järjestusi ning nende kohta käivaid metaandmeid, nagu näiteks proovivõtmise paik, peremeestaime liik, kasvukohatüüp jne. Andmebaas sisaldab SSU, ITS ja LSU piirkondade DNA järjestusi, mis on juba teadusartiklites avaldatud. AM seente uurimiseks kasutatakse sageli SSU markerit, sest see piirkond võimaldab eristada suurt osa AM seente liike ning on hästi kasutatav fülogeneesianalüüsiks informatiivse piirkonna pikkuse ning varieeruvuse määra tõttu. AM seente SSU markerpiirkonna järjestused on fülogeneesianalüüsi alusel rühmitatud umbes liigi tasemel taksoniteks, mida nimetatakse virtuaaltaksoniteks (VT) ja mis toimivad elurikkuse kirjeldamisel klassikaliste liikide analoogina. Maarj*AM* on ekspertide poolt käsitsi kureeritav andmebaas, mis tagab andmete kõrge kvaliteedi ning seetõttu on kasutatav referents(võrdlus)andmestikuna. Uute DNA järjestuste võrdlus, näiteks BLAST algoritmi abil, selle andmebaasi vastu võimaldab kergesti määrata AM seente järjestusi, mis on DNA järjestamise teel tuvastatud näiteks eostest, taimejuurtest ja mullaproovidest.

Ükski andmebaas pole täielik, seega ei pruugi uute proovide analüüsimisel leida liigi tasemel vastet kõikidele järjestustele Maarj*AM* andmebaasist. Täpse vaste puudumisel on võimalik selliseid järjestusi määrata kasutades suurimat nukleotiidsete järjestuste avalikku andmebaasi, milleks on geenipankade konsortsium International Nucleotide Sequence Database Collaboration (INSDC). INSDC sisaldab nii teadusajakirjades avaldatud kui ka avaldamata DNA järjestusi koos metaandmetega. Antud andmebaasi kasutatakse, et tuvastada uute järjestuste taksonoomiline kuuluvus olenevalt annoteeritud täpsusele: madalaimaks tasemeks on liik ja kõrgeim taksonoomiline tase on riik (näiteks: seened). Leides täiendavaid määranguid meile huvipakkuvate organismide rühmades, saab uusi järjestusi fülogeneetilise analüüsi abil lisada olemasolevasse referentsandmebaasi ja seda järjest täielikumaks muuta. See võimaldab täiendada Maarj*AM* andmebaasi looduses esinevate, kuid andmebaasis seni esindamata liikidega, mis aitavad järgmistes suuremahulistes sekveneerimisandmetes varasemast täielikumalt ja täpsemalt määrata otsitavaid organisme.

Mitmed uuringud on hinnanud INSDC andmebaasi kirjete kvaliteeti ja tuvastanud kitsaskohti. See andmebaaside kogum sisaldab DNA järjestusandmeid, mis võivad olla kimäärsed (järjestused, mis on kombineeritud kahest või enamast organismist pärit järjestusest), madala kvaliteediga (sisaldab määramata nukleotiide, N), valesti määratud või ebapiisavalt annoteeritud. Taolised ebatäpsused on paratamatud avalike andmekogude puhul, mida täielikult ekspertide poolt kureerida ei ole võimalik või otstarbekas, seetõttu luuakse eriotstarbelisi kureeritud

andmebaase, mis on väiksemad ja kõrge kvaliteediga, nagu ka näiteks INSDC üks osa-andmebaase RefSeq. AM seente määramiseks SSU geenipiirkonnas kasutatavate AM seenespetsiifiliste praimeritega on järjestuste sekveneerimisel üheks kõrvalproduktiks muud juurtes ja mullas elutsevate organismide järjestused (näiteks muud seened, mullaloomad või vetikad), mida on võimalik määrata INSDC andmebaasiga võrdluste abil. Kuigi need andmed, niinimetatud mittespetsiifilised järjestused, võimaldavad saada üldisema pildi mullas ja taimejuurtes esinevatest elustikest, on nende organismide elurikkuse täpsemateks uuringuteks vaja kasutada kogu huvipakkuvat organismirühma tuvastavaid praimereid ning järjestuste määramisel jälgida, millised andmebaasid on nende organismide kohta olemas.

Võrdlusandmebaasi olemasolule ning omadustele on DNA järjestuste määramisel oluline protsess nii-öelda tooreste sekveneerimisandmete esmane puhastamine ja filtreerimine. Järjestused jaotatakse proovidesse vöötkoodi (*barcode*) alusel, seejärel puhastatakse filtreeritud järjestused praimeri või praimerite paari alusel, mida on kasutatud antud markerpiirkonna amplifitseerimiseks. Järjestuste filtreerimise parameetrid aitavad vähendada sekveneerimisvigadega andmete sattumist järgnevatesse analüüsi etappidesse, sest valitakse kõrge kvaliteedihinnangu saanud järjestusi, mille vöötkood ja praimerite järjestused on kattuvad kasutaja poolt etteantud järjestustega. Probleemne on ka järjestuste paigutamine valedesse proovidesse (*barcode-leakage*), mille lahendusena rakendatakse proove eristatavate vöötkoodide koostamisel põhimõtet, et iga vöötkood peab erinema teisest koodist vähemalt kahe nukleotiidi vea või vahetuse võrra. Illumina sekveneerimisplatvormi puhul kasutatakse paarisjärjestusi (eesmine ja tagumine järjestus: R1 ja R2), et vähendada antud probleemi, sest vöötkoodide paar peab sobituma mõlema paarisjärjestusega korraga. Sellisel juhul peaks toimuma vähemalt 4 nukleotiidi vahetus, et järjestuste paarid valesse proovi grupeeritaks. Paarisjärjestused toimivad põhimõttel, et eesmise ja tagumise järjestuse ülekattumise piirkond asub järjestuse lõpus, mis on üldjuhul madalama kvaliteediga sekveneerimisplatvormidest tulenevate iseärasuste tõttu ja võib sisalda järjestusvigu. Paarisjärjestuste rakendamine aitab programmidel, mis neid kombineerivad, leida ühelt või teiselt järjestuselt kõrgema kvaliteediga nukleotiidi ja seeläbi parandada järjestusvigu.

Amplifitseeritud järjestuste puhul on üheks probleemiks andmete tõesus, sest PCR amplifikatsiooni protsessi käigus võib tekkida DNA järjestusi, mis koosnevad algselt mitmest organismist pärinevate järjestuste osadest (eelmainitud kimäärsed järjestused). Kimäärsed järjestused võivad oluliselt suurendada tuvastatud liikide arvu hinnangut, kui kimäärseid järjestusi peetakse uuteks liikideks. Mitmeid programme on loodud kimäärsete järjestuste äratundmiseks ning eemaldamiseks. Tuvastamiseks jaotatakse järjestused väiksemateks osadeks ning uuritakse neid vastu kimäärivaba võrdlusandmebaasi, mille tulemusel leitakse igale osale andmebaasist lähimad kandidaadid. Kui neid kandidaate on mitu, erinevad osade vahel ja need pärinevad erinevatest organismidest, antakse neile usaldusväärtused, mille alusel programmis kirjeldatud algoritm püüab liigitada järjestusi potentsiaalselt kimäärseteks või nende vabadeks. Kimääride eemaldamisel saab puhastatud

andmed määrata liikidesse mitmel viisil. Üheks enimkasutatavaks meetodiks on järjestuste klasterdamine (rühmitamine) teatud sarnasusprotsendi alusel taksonoomilisteks üksusteks, mida on võimalik määrata liigi või kõrgemal taksonoomilisel tasemel võrreldes vastava andmebaasi määrangute järjestustega. Klasterdamine on tõhus sellistele markerpiirkondadele ja organismidele, mille võrdlusandmestik on ebatäielik ja ei hõlma veel piisavat hulka organisme antud rühmast. Teiseks võimaluseks on võrrelda iga järjestust individuaalselt vastu võrdlusandmebaasi kasutades selleks BLAST algoritmi, mis joondab sisendjärjestused vastu referentsjärjestusi ja annab tulemuseks parima sobiliku kandidaadi koos identsuse ja joonduse (kahe järjestuse kõrvutuse) parameetritega. Näitajaid saab kasutada leitud vastete filtreerimiseks. Näiteks saab filtreerimiseks joonduse pikkuse määra kasutades tuvastada ja seejärel eemaldada järjestusi, mis võisid läbida kimäärikontrolli, aga joonduvad ainult järjestuse osalise pikkusega vastu andmebaasijärjestust.

Eelkirjeldatud DNA järjestuste määramise etapid on iga organismirühma ja DNA järjestamise platvormi puhul mõnevõrra erinevad. Bioinformaatilised tööriistad vajavad kohandamist nii organismirühmale, selle markerpiirkonnale kui DNA järjestamise tehnoloogiale. Minu doktoritöö peamised eesmärgid on järgmised:

1) Uurida erinevate markerpiirkondade varieeruvust AM seente liikide piires ja liikide vahel ning tuvastada varieeruvad piirkonnad, mis on liikide määramisel otstarbekad.
2) Võrrelda erinevate DNA sekveneerimisplatvormidega saadud järjestuseandmete tehnilisi erinevusi ning hinnata, kui edukalt suudavad need platvormid tuvastada AM seeni keskkonnaproovidest.
3) Võrrelda SSU ja ITS markerpiirkondade alusel AM seente liigilist koosseisu ja liigirikkust, hinnata selle muutust erinevate keskkonnatingimuste mõjul.
4) Koostada tarkvaraline töövahend, mis hõlbustaks AM seente DNA-põhist määramist.

Artikkel **I** kirjeldab kahe AM seeneliigi (*Rhizophagus irregularis* ja *Gigaspora margarita*) varieeruvust liigisiseselt ja liikide vahel kolme ribosomaalse DNA markerpiirkonna osas. Selleks eraldati eostest DNA, amplifitseeriti markerpiirkonnad PCR teel, kloneeriti ja sekveneeriti kloonid Sanger sekveneerimisplatvormiga. Töös leiti, et kõik uuritud markerpiirkonnad võimaldavad eristada antud kahte AM seent ja omavad liikide eristamiseks piisavalt varieeruvust (*barcode gap*). Teisisõnu, liigisisene varieeruvus (*intraspecific variation*) on madalam kui liikide vaheline varieeruvus (*interspecific variation*). Varieeruvus oli kõige suurem ITS piirkonna järjestustel ja kõige madalam SSU markeril. Samas, SSU hõlmas endas alampiirkonda, mis varieerus rohkem kui ITS piirkond, tehes sellest sobiva kandidaadi AM seente määramiseks. Lisaks on SSU piirkonna puhul AM seened liigi tasemel eristatavad ja mittevarieeruvate külgmiste piirkondade tõttu on seda hea kasutada fülogeneesianalüüsis, sest erinevate AM seenerühmade järjestused on hõlpsalt joondatavad ja võrreldavad. ITS piirkonna puhul ei ole see sugukonnast

kõrgemate taksonite puhul võmalik, sest DNA järjestuste varieeruvus on liiga suur nende sisukaks joondamiseks.

Artikkel **II** uuris tollel hetkel laialt kasutatava 454 sekveneerimisplatvormiga saadud järjestuseandmete erinevusi uuema sekveneerimisplatvormiga Illumina MiSeq saadud DNA järjestustega ja kuivõrd tehnoloogilised erinevused mõjutavad ökoloogilisi järeldusi. Lisaks uuriti SSU markeri NS31-AML2 praimeripaari amplikoni DNA järjestuste varieeruvust kasutades selleks Maarj*AM* andmebaasis olevaid avaldatud järjestusi. Leidsime, et pärast NS31 praimeri asukohta esineb varieeruv piirkond 70. ja 300. nukleotiidipositsiooni vahel ning ülejäänud osa amplikoni järjestusest on AM seente piires vähe varieeruv. Illumina MiSeq sekveneerimisplatvormiga oli tollal võimalik järjestada 2×250 aluspaari pikkuseid järjestusi, mis ei ole piisav AM seente SSU täisamplikoni kasutamiseks NS31 ja AML2 praimerite vahel. Illumina platvormil sekveneerimiseks kasutati seetõttu tagmentatsiooni, kus järjestused algavad amplikonil juhuslikult, mitte praimeripositsioonilt. Tagmentatsioonifragmentide järjestuste määramine ilma põhjaliku filtreerimiseta andis tulemuseks kunstlikult kõrge AM seente liigirikkuse hinnangu võrreldes samadest proovidest 454 järjestamisel saadud tulemustega. Selgus, et tagmentatsioonil saadud Illumina järjestused, mis asusid SSU amplikoni mittevarieeruvas piirkonnas, ei olnud BLAST algoritmiga korrektselt liigi tasemel tuvastatavad, põhjustades liigirikkuse hinnangu suure tõusu. Liigirikkuse andmete usaldusväärseks kasutamiseks on tähtis järjestusi õigesti filtreerida ja olla teadlik amplikoni iseärasustest, sh varieeruvusest. Kasutades filtreerimisparameetreid leidsime, et 454 ja Illumina järjestuste põhiselt saadud AM seente liigirikkus oli sarnane ja enim levinud VT olid samad. Harva esinevad VT erinesid mõlemal platvormil, mille üheks põhjuseks võivad olla tehnoloogilised iseärasused või ka juhus. Lisaks uuriti, kas AM seeni on võimalik tuvastada ka taimejuurte metagenoomse DNA hulgast. Sellise metoodika puhul pole tarvis PCR amplifitseerimist kasutada, mis vähendaks kimäärsete järjestuste tekkimise võimalusi. Selgus, et AM seente osakaal kogu sekveneeritud DNA hulgas on väike ja enamus tuvastatud eukarüootsest pärilikkusainest kujutas endast peremeestaime järjestusi. Lisaks leidsime, et Illumina MiSeq platvormiga saadav suurem järjestuste arv proovi kohta ei suurendanud leitud AM seente liigirikkust võrreldes 454 sekveneerimisega, kus saadakse proovist väiksem arv järjestusi. See tulemus viitab, et 454 järjestamisega on saavutatud piisav sekveneerimis-sügavus AM seente uurimiseks ning täiendavat sekveneerimissügavuse kasvu elurikkuse hinnangute parandamiseks pole tarvis.

Praimereid uuendatakse ja täiendatakse pidevalt, et nad võimaldaks tuvastada paljusid organismigruppe ja parandada pilti elurikkuse mustrite kohta. Artikkel **III** uuris, kas AM seente määramiseks kasutatava SSU piirkonna järjestamisel tuvastatav AM seente liigirikkus ja elurikkuse ökoloogilised mustrid on sarnased kõigi seente määramiseks kasutatava ITS piirkonna järjestamisel saadavate andmetega. Selleks kasutati kahte katset, kus ühes mõjutati taimi niiske (30% vett) ja kuiva (10% vett) kasvukeskkonnaga ja teises võrreldi AM seente koosluste muutumist hooajaliselt kolmel erineval kuul: aprill, juuli ja september.

Leidsime, et enamlevinud AM seeni on võimalik ka ITS piirkonda seenespetsiifiliste parimeritega järjestades tuvastada, kuigi osa liike jääb ka tuvastamata. Lisaks, mõlema markerpiirkonnaga saadud andmetes oli näha sarnaseid muutuseid AM seente kooslustes sõltuvalt keskkonnatingimustest ning proovivõtuajast. Lisaks on laiema spetsiifilisusega praimeritega ITS piirkonda järjestades võimalik tuvastada palju erinevaid seeni, kasutades liikide määramiseks UNITE andmebaasi. Olenevalt teaduslikust uurimisküsimustest võib olla otstarbekas kasutada ITS piirkonna järjestamist kõiki seeni kindlakstegeva metoodikaga ka AM seente koosluste uurimiseks, kui liigirikkuse hinnangu täpsus pole keskse tähtsusega. Lisaks võimaldab antud ITS-põhine lähenemine kasutada lühemat järjestuse pikkust võrreldes AM seente-spetsiifilise SSU järjestamisega ja kasutada ainult eesmisi järjestusi, muutes sekveneerimise läbiviimise odavamaks.

Viimane käsikiri (**IV**) esitab koostatud bioinformaatilise abivahendi gDAT (*graphical downstream analysis tool*), mis võimaldab lihtsalt ja mugavalt hallata ja analüüsida pärilikkusaine järjestamisel põhinevat elurikkuse informatsiooni. Lisaks seab antud bioinformaatiline abivahend ette kindla raamistiku ja töövoo, mille järgimine võimaldab võrrelda omavahel erinevaid läbitöödeldud andmestikke ning vähendab kasutajate poolt tekitatud sisestusvigu andes ette soovituslikud ja eelseadistatud parameetrid. Testandmestikuna kasutati Illumina MiSeq järjestusi, mis olid saadud kahe erineva markerpiirkonnaga (SSU ja ITS) ja kolme erineva praimeripaariga, et tuvastada AM seente (SSU), kõigi seente (ITS) ja kõigi eukarüootide (SSU) elurikkust mullas. Programmi valideerimiseks mõõdeti iga bioinformaatilise etapi kohta protsessori, mälu ja kõvaketta kasutust ning ajakulu uuritud andmestiku põhjal. Peale selle toodi välja iga etapi läbinud järjestuste hulgad ning kõvaketta ja mälu kogukasutus ning kogu ajakulu. Töö näitas, et Illumina MiSeq järjestuseandmestikke on võimalik ühe ööpäeva jooksul analüüsida, kasutades selleks tavapärast üldkasutusel saadaolevat arvutit ning ei ole vajadust kasutada suure jõudlusega klasterarvuti ressursse.

Kokkuvõtvalt, doktoritöö üheks peamiseks tulemuseks on bioinformaatiline abivahend gDAT, mis hõlbustab AM seente DNA-põhiseid uuringuid keskkonnast SSU markeri põhiselt. Lisaks on seda töövahendit võimalik erinevate võrdlusandmebaaside kasutamisel rakendada ka teiste markerpiirkondade ja organismirühmade andmestikega. Töö pakub välja võimalusi ja lahendusi, milliseid genoomipiirkondi on otstarbekas kasutada, et kindlaks teha ja määrata AM seeni keskkonnaproovidest. Võrdlesin erinevate markerpiirkondade efektiivsust AM seente määramisel ning uurisin lähemalt geeni- või markerpiirkondade varieeruvust. Doktoritöö tulemusena selgus, et AM seente määramiseks sobib hästi geenipiirkond SSU, mis on konstantse järjestuspikkusega, hõlmab hüpervarieeruvat ja seega informatiivset piirkonda järjestuse keskel ja mittevarieeruvaid piirkondi järjestuse eesmises ja tagumises osas. Need tunnused teevad SSU geenipiirkonna järjestused lihtsasti joondatavaks erinevate organismirühmade vahel fülogeneesianalüüsides. Samuti on võimalik järjestusi liigiliselt eristada BLAST algoritmi abil, sest antud piirkonnas on liigisisene varieeruvus väiksem kui liikide vahel. Lisaks võimaldab SSU geenipiirkond edukalt sekveneerida paljusid AM seeneliike.

ACKNOWLEDGEMENTS

My sincerest gratitude goes to Dr. Maarja Öpik and Prof. Peter Young. Thank you for the comprehensive supervision, and sharing your knowledge and expertise during the years of my doctoral studies. I am thankful to all of the co-authors, and especially to John Davison for statistical advice. I also thank Prof. Martin Zobel and Dr. Mari Moora for giving the opportunity to work with plant ecologists and to allow me to study the new, interesting and developing field of bioinformatics. I am most grateful to all of my colleagues from the plant ecology workgroup for creating a friendly working atmosphere, providing stimulating discussions and sharing great ideas. I also thank my parents and family, my wife and son, for making my passage through the hard times easier and for pushing me to finish the work I started.

PUBLICATIONS

ELULOOKIRJELDUS

Nimi: Martti Vasar
Sünniaeg: 12.02.1987
Kodakondsus: Eesti
E-mail: martti.vasar@ut.ee

Hariduskäik:
2003–2006   Tartu Kommertsgümnaasium
2006–2010   Tartu Ülikool, B.Sc. infotehnoloogia
2010–2012   Tartu Ülikool, M.Sc. infotehnoloogia
2014–…   Tartu Ülikool, botaanika ja ökoloogia doktorantuur

Konverentsiettekanded:
Vasar M, Öpik M. Workflow to analyse arbuscular mycorrhizal fungi in plant roots and soil. (suuline ettekanne) Doctoral Student Conference of the Departments of Botany and Zoology, University of Tartu, 17.–18.11.2016, Kubija, Estonia.

Vasar M, Öpik M. Comparison of 454 and Illumina-sequencing methods to study arbuscular mycorrhizal fungal community diversity. (poster ettekanne) British Ecological Society Annual Meeting 2016, 11.–14.12.2016, Liverpool, Inglismaa.

Kursused:
2017   Kursus „Bioinformatics for Adaptation Genomics" 5.–11.02.2017, Weggis, Šveits.

Publikatsioonid:
Gavito ME, Leyva-Morales R, Vega-Peña EV, Arita H, Jairus T, **Vasar M**, Öpik M. 2019. Local-scale spatial diversity patterns of ectomycorrhizal fungal communities in a subtropical pine-oak forest. *Fungal Ecology*, 42:100860.

Noreika N, Helm A, Öpik M, Jairus T, **Vasar M**, Reier Ü, Kook E, Riibak K, Kasari L, Tullus H, Tullus T. 2019 Forest biomass, soil and biodiversity relationships originate from biogeographic affinity and direct ecological effects. *Oikos*, 128(11):1653–1665.

Peña-Venegas CP, Kuyper TW, Davison J, Jairus T, **Vasar M**, Stomph TJ, Struik PC, Öpik M. 2019. Distinct arbuscular mycorrhizal fungal communities associate with different manioc landraces and Amazonian soils. *Mycorrhiza*, 29(3):263–275.

Träger S, Öpik M, **Vasar M**, Wilson SD. 2019. Belowground plant parts are crucial for comprehensively estimating total plant richness in herbaceous and woody habitats. *Ecology*, 100(2):e02575.

Sepp S, Davison J, Jairus T, **Vasar M**, Moora M, Zobel M, Öpik M. 2019. Non-random association patterns in a plant-mycorrhizal fungal network reveal host-symbiont specificity. *Molecular Ecology*, 28(2):365–378.

Faggioli VS, Cabello MN, Grilli G, **Vasar M**, Covacevich F, Öpik M. 2019. Root colonizing and soil borne communities of arbuscular mycorrhizal fungi differ among soybean fields with contrasting historical land use. *Agriculture, Ecosystems and Environment*, 269:174–182.

Jiang S, Liu Y, Luo J, Qin M, Johnson NC, Öpik M, Vasar M, Chai Y, Zhou X, Mao L, Du G. 2018. Dynamics of arbuscular mycorrhizal fungal community structure and functioning along a nitrogen enrichment gradient in an alpine meadow ecosystem. *New Phytologist*, 220(4):1222–1235.

Lekberg Y, **Vasar M**, Bullington LS, Sepp S, Antunes PM, Bunn R, Larkin BG, Öpik M. 2018. More bang for the buck? Can arbuscular mycorrhizal fungal communities be characterized adequately alongside other fungi using general fungal primers? *New Phytologist*, 220(4):971–976.

Neuenkamp L, Moora M, Öpik M, Davison J, Gerz M, Männistö M, Jairus T, **Vasar M**, Zobel M. 2018. The role of plant mycorrhizal type and status in modulating the relationship between plant and arbuscular mycorrhizal fungal communities. *New Phytologist*, 220(4):1236–1247.

Davison J, Moora M, Opik M, Ainsaar L, Ducousso M, Hiiesalu I, Jairus T, Johnson N, Jourand P, Kalamees R, Koorem K, Meyers J, Pussa K, Reier Ul, Partel M, Semchenko M, Traveset A, **Vasar M**, Zobel M. 2018. Microbial island biogeography: isolation shapes the life history characteristics but not diversity of root-symbiotic fungal communities. *The ISME Journal*, 12(9): 2211–2224.

García de León D, Davison J, Öpik M, Hiiesalu I, Jairus T, Koorem K, Liu Y, Phosri C, Sepp S, **Vasar M**, Zobel M. 2018. Anthropogenic disturbance equalizes diversity levels in arbuscular mycorrhizal fungal communities. *Global Change Biology*, 24(6):2649–2659.

García de León D, Neuenkamp L, Moora M, Öpik M, Davison J, Peña-Venegas CP, **Vasar M**, Jairus T, Zobel M. 2018. Arbuscular mycorrhizal fungal communities in tropical rain forest are resilient to slash-and-burn agriculture. Journal of Tropical Ecology, 34(3):186–199.

Sepp S, Jairus T, **Vasar M**, Zobel M, Öpik M. 2018. Effects of land use on arbuscular mycorrhizal fungal communities in Estonia. *Mycorrhiza*, 28(3):259–68.

García de León D, Cantero JJ, Moora M, Öpik M, Davison J, **Vasar M**, Jairus T, Zobel M. 2018. Soybean cultivation supports a diverse arbuscular mycorrhizal fungal community in central Argentina. *Applied Soil Ecology*, 124:289–297.

Lopez-Garcia A, Varela-Cervero S, **Vasar M**, Öpik M, Barea JM, Azcon-Aguilar C. 2017. Plant traits determine the phylogenetic structure of arbuscular mycorrhizal fungal communities. *Molecular Ecology*, 26(24):6948–6959.

Vasar M, Andreson R, Davison J, Jairus T, Moora M, Remm M, Young JPW, Zobel M, Öpik M. 2017. Increased sequencing depth does not increase captured diversity of arbuscular mycorrhizal fungi. *Mycorrhiza*, 27(8):761–773.

Koorem K, Tulva I, Davison J, Jairus T, Öpik M, **Vasar M**, Zobel M, Moora M. 2017. Arbuscular mycorrhizal fungal communities in forest plant roots are simultaneously shaped by host characteristics and canopy-mediated light availability. *Plant and Soil*, 410(1–2):259–271.

Ciccolini V, Ercoli L, Davison J, **Vasar M**, Öpik M, Pellegrino E. 2016. Land-use intensity and host plant simultaneously shape the composition of arbuscular mycorrhizal fungal communities in a Mediterranean drained peatland. *FEMS Microbiology Ecology*, 92(12):fiw186.

García de León D, Moora M, Öpik M, Jairus T, Neuenkamp L, **Vasar M**, Bueno CG, Gerz M, Davison J, Zobel M. 2016. Dispersal of arbuscular mycorrhizal fungi and plants during succession. *Acta Oecologica*, 77:128–135.

Hermann L, Lesueur D, Bräu L, Davison J, Jairus T, Robain H, Robin A, **Vasar M**, Wiriyakitnateekul W, Öpik M. 2016. Diversity of root-associated arbuscular mycorrhizal fungal communities in a rubber tree plantation chronosequence in Northeast Thailand. *Mycorrhiza*, 26(8):863–877.

Moora M, Öpik M, Davison J, Jairus T, **Vasar M**, Zobel M, Eckstein R. 2016. AM fungal communities inhabiting the roots of submerged aquatic plant Lobelia dortmanna are diverse and include a high proportion of novel taxa. *Mycorrhiza*, 26(7):735–745.

García de León D, Moora M, Öpik M, Neuenkamp L, Gerz M, Jairus T, **Vasar M**, Bueno C, Davison J, Zobel M. 2016. Symbiont dynamics during ecosystem succession: co-occurring plant and arbuscular mycorrhizal fungal communities. *FEMS Microbiology Ecology*, 92(7):fiw097.

Thiéry O, **Vasar M**, Jairus T, Davison J, Roux C, Kivistik P, Metspalu A, Milani L, Saks Ü, Moora M, Zobel M, Öpik M. 2016. Sequence variation in nuclear ribosomal small subunit, internal transcribed spacer and large subunit regions of Rhizophagus irregularis and Gigaspora margarita is high and isolate-dependent. *Molecular Ecology*, 25(12):2816–2832.

Gazol A, Zobel M, Jose Cantero J, Davison J, Esler KJ, Jairus T, Öpik M, **Vasar M**, Moora M. 2016. Impact of alien pines on local arbuscular mycorrhizal fungal communities-evidence from two continents. *FEMS Microbiology Ecology*, 92(6):fiw073.

Davison J, Moora M, Jairus T, **Vasar M**, Öpik M, Zobel M. 2016. Hierarchical assembly rules in arbuscular mycorrhizal (AM) fungal communities. *Soil Biology and Biochemistry*, 97:63–70.

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, Johnson NC, Kane A, Koorem K, Kochar M, Ndiaye C, Pärtel M, Reier Ü, Saks Ü, Singh R, **Vasar M**, Zobel M. 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science*, 349(6251):970–973.

Varela-Cervero S, **Vasar M**, Davison J, Barea JM, Öpik M, Azcón-Aguilar C. 2015. The composition of arbuscular mycorrhizal fungal communities differs among the roots, spores and extraradical mycelia associated with five Mediterranean plant species. *Environmental Microbiology*, 17(8):2882–2895.

Grilli G, Urcelay C, Galetto L, Davison J, **Vasar M**, Saks Ü, Jairus T, Öpik M. 2015. The composition of arbuscular mycorrhizal fungal communities in the roots of a ruderal forb is not related to the forest fragmentation process. *Environmental microbiology*, 17(8):2709–2720.

Moora M, Davison J, Öpik M, Metsis M, Saks Ü, Jairus T, **Vasar M**, Zobel M. 2014. Anthropogenic land use shapes the composition and phylogenetic structure of soil arbuscular mycorrhizal fungal communities. *FEMS microbiology ecology*, 90(3):609–621.

Hiiesalu I, Pärtel M, Davison J, Gerhold P, Metsis M, Moora M, Öpik M, **Vasar M**, Zobel M, Wilson S. 2014. Species richness of arbuscular mycorrhizal fungi: associations with grassland plant richness and biomass. *New Phytologist*, 203(1):233–244.

Saks Ü, Davison J, Öpik M, **Vasar M**, Moora M, Zobel M. 2014. Root-colonizing and soil-borne communities of arbuscular mycorrhizal fungi in a temperate forest understorey. *Botany*, 92(4):277–285.

Öpik M, Zobel M, Cantero J, Davison J, Facelli J, Hiiesalu I, Jairus T, Kalwij J, Koorem K, Leal M, Liira J, Metsis M, Neshataeva V, Paal J, Phosri C, Põlme S, Reier Ü, Saks Ü, Schimann H, Thiery O, **Vasar M**, Moora M. 2013. Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza*, 23(5):411–430.

Thiéry O, Moora M, **Vasar M**, Zobel M, Öpik M. 2012. Inter- and intrasporal nuclear ribosomal gene sequence variation within one isolate of arbuscular mycorrhizal fungus, Diversispora sp. *Symbiosis*, 58(1–3):135–147.

Vasar M, Srirama SN, Dumas M. 2012. Framework for Monitoring and Testing Web Application Scalability on the Cloud. In *Proceedings of the WICSA/ ECSA 2012 Companion Volume* (pp. 53–60).

Hiiesalu I, Öpik M, Metsis M, Davison J, **Vasar M**, Moora M, Zobel M, Wilson S, Pärtel M. 2012. Plant species richness belowground: higher richness and new patterns revealed by next generation sequencing. *Molecular Ecology*, 21(8):2004–2016.

Davison J, Öpik M, Zobel M, **Vasar M**, Metsis M, Moora M. 2012. Communities of arbuscular mycorrhizal fungi detected in forest soil are spatially heterogeneous but do not vary throughout the growing season. *PLoS One*, 7(8): e41938.

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 micro-organisms 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. **Krista Kaasik**. 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 intra-cellular 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 cellobio-hydrolase. 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 related-ness, 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 geo-graphic 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.

340. **Märt Roosaare.** *K*-mer based methods for the identification of bacteria and plasmids. Tartu, 2018, 117 p.

341. **Maria Abakumova.** The relationship between competitive behaviour and the frequency and identity of neighbours in temperate grassland plants. Tartu, 2018, 104 p.

342. **Margus Vilbas.** Biotic interactions affecting habitat use of myrmecophilous butterflies in Northern Europe. Tartu, 2018, 142 p.

343. **Liina Kinkar**. Global patterns of genetic diversity and phylogeography of *Echinococcus granulosus* sensu stricto – a tapeworm species of significant public health concern. Tartu, 2018, 147 p.

344. **Teivi Laurimäe.** Taxonomy and genetic diversity of zoonotic tapeworms in the species complex of *Echinococcus granulosus* sensu lato. Tartu, 2018, 143 p.

345. **Tatjana Jatsenko.** Role of translesion DNA polymerases in mutagenesis and DNA damage tolerance in Pseudomonads. Tartu, 2018, 216 p.

346. **Katrin Viigand.** Utilization of α-glucosidic sugars by *Ogataea* (*Hansenula*) *polymorpha.* Tartu, 2018, 148 p.

347. **Andres Ainelo.** Physiological effects of the *Pseudomonas putida* toxin grat. Tartu, 2018, 146 p.

348. **Killu Timm**. Effects of two genes (DRD4 and SERT) on great tit (*Parus major*) behaviour and reproductive traits. Tartu, 2018, 117 p.

349. **Petr Kohout**. Ecology of ericoid mycorrhizal fungi. Tartu, 2018, 184 p.

350. **Gristin Rohula-Okunev**. Effects of endogenous and environmental factors on night-time water flux in deciduous woody tree species. Tartu, 2018, 184 p.

351. **Jane Oja.** Temporal and spatial patterns of orchid mycorrhizal fungi in forest and grassland ecosystems. Tartu, 2018, 102 p.

352. **Janek Urvik.** Multidimensionality of aging in a long-lived seabird. Tartu, 2018, 135 p.

353. **Lisanna Schmidt**. Phenotypic and genetic differentiation in the hybridizing species pair *Carex flava* and *C. viridula* in geographically different regions. Tartu, 2018, 133 p.

354. **Monika Karmin.** Perspectives from human Y chromosome – phylogeny, population dynamics and founder events. Tartu, 2018, 168 p.

355. **Maris Alver.** Value of genomics for atherosclerotic cardiovascular disease risk prediction. Tartu, 2019, 148 p.

356. **Lehti Saag**. The prehistory of Estonia from a genetic perspective: new insights from ancient DNA. Tartu, 2019, 171 p.

357. **Mari-Liis Viljur**. Local and landscape effects on butterfly assemblages in managed forests. Tartu, 2019, 115 p.

358. **Ivan Kisly.** The pleiotropic functions of ribosomal proteins eL19 and eL24 in the budding yeast ribosome. Tartu, 2019, 170 p.

359. **Mikk Puustusmaa**. On the origin of papillomavirus proteins. Tartu, 2019, 152 p.

360. **Anneliis Peterson**. Benthic biodiversity in the north-eastern Baltic Sea: mapping methods, spatial patterns, and relations to environmental gradients. Tartu, 2019, 159 p.

361. **Erwan Pennarun.** Meandering along the mtDNA phylogeny; causerie and digression about what it can tell us about human migrations. Tartu, 2019, 162 p.

362. **Karin Ernits**. Levansucrase Lsc3 and endo-levanase BT1760: characterization and application for the synthesis of novel prebiotics. Tartu, 2019, 217 p.

363. **Sille Holm**. Comparative ecology of geometrid moths: in search of contrasts between a temperate and a tropical forest. Tartu, 2019, 135 p.

364. **Anne-Mai Ilumäe.** Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians. Tartu, 2019, 172 p.

365. **Anu Lepik**. Plant competitive behaviour: relationships with functional traits and soil processes. Tartu, 2019, 152 p.

366. **Kunter Tätte.** Towards an integrated view of escape decisions in birds under variable levels of predation risk. Tartu, 2020, 172 p.

367. **Kaarin Parts.** The impact of climate change on fine roots and root-associated microbial communities in birch and spruce forests. Tartu, 2020, 143 p.

368. **Viktorija Kukuškina.** Understanding the mechanisms of endometrial receptivity through integration of 'omics' data layers. Tartu, 2020, 169 p.