DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
459

ALI HAKIMZADEH

Long-read metabarcoding: from available tools to reference databases

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
459

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
459

ALI HAKIMZADEH

Long-read metabarcoding: from available tools to reference databases

UNIVERSITY OF TARTU
Press
1632

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 Botany and Mycology at the University of Tartu on 12.01.2026 by the Scientific Council of the Institute of Ecology and Earth Sciences University of Tartu.

Supervisors: Dr. Sten Anslan, University of Jyväskylä, Finland

Prof. Leho Tedersoo, University of Tartu, Estonia

Opponent: Dr. Owen S. Wangensteen, University of Barcelona, Spain

Commencement: Oecologicum (J. Liivi 2, Tartu), room 127, on the 13.02.2026 at 10.15 a.m.

Publication of this thesis is granted by the Institute of Ecology and Earth Sciences, University of Tartu

ISSN 1024-6479 (print)
ISBN 978-9908-57-128-7 (print)
ISSN 2806-2140 (pdf)
ISBN 978-9908-57-129-4 (pdf)

University of Tartu Press
www.tyk.ee

CONTENTS

LIST OF PUBLICATIONS

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

I. **Hakimzadeh, A.,** Asbun, A. A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Durling, M. B., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., Palmer, J. M., Pascal, G., Porter, T. M., Straub, D., Vasar, M., Větrovský, T., Zafeiropoulos, H., Anslan, S. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. Molecular Ecology Resources, 24(5), e13847.

II. Tedersoo, L., Hosseyni Moghaddam, M. S., Mikryukov, V., **Hakimzadeh, A.,** Bahram, M., Nilsson, R. H., Yatsiuk, I., Geisen, S., Schwelm, A., Piwosz, K., Prous, M., Sildever, S., Chmolowska, D., Rueckert, S., Skaloud, P., Laas, P., Tines, M., Jung, J.-H., Choi, J. H., Alkahtani, S., Anslan, S. (2024). EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. Database, baae043.

III. **Hakimzadeh, A.,** Mikryukov, V., Metsoja, M., Tedersoo, L., Anslan, S. (2025) Are we throwing away good data? Evaluation of chimera detection algorithms on long-read amplicons reveals high false-positive rates across algorithms. PeerJ, 13, e20456.

Author's contribution ('*' denotes a moderate contribution, '**' denotes a high contribution '***' denotes a leading role).

	I	II	III
Original idea	**	–	***
Study design	***	*	***
Data collection	***	**	***
Analysis and interpretation	***	**	***
Manuscript writing	***	*	***

LIST OF ABBREVIATIONS

CLI – command line interface
COI – cytochrome oxidase subunit I
eDNA – Environmental DNA
ESV – exact sequence variance
GUI – graphical user interface
HSP – high-scoring segment pair
HTS – high-throughput sequencing
INSDC – International Nucleotide Sequence Database Collaboration
ITS – Internal Transcribed Spacer
LSU – large subunit of ribosomal RNA
NMDS – non-metric multidimensional scaling
NUMTs – nuclear mitochondrial pseudogenes
OTU – operational taxonomic unit
PacBio – Pacific Biosciences
PCR – polymerase chain reaction
rRNA – ribosomal ribonucleic acid
SSU – small subunit of ribosomal RNA

LIST OF TERMS AND DEFINITIONS

BLAST – Basic Local Alignment Search Tool, an algorithm that compares a query sequence against a database, identifies local similarity regions, and reports key alignment statistics such as length, matches, mismatches, and gaps.

DNA Barcode – A short, standardized DNA sequence to identify and differentiate species.

Environmental DNA (eDNA) – Genetic material extracted from environmental samples like water or soil.

Metabarcoding – A PCR-based method for species identification from a sample containing DNA from more than one organism.

1. INTRODUCTION

The ongoing loss of the Earth's biodiversity is one of the most pressing challenges of the century (Thomsen & Willerslev, 2015). However, tracking it by relying on conventional approaches such as morpho-taxonomy for species monitoring can be challenging, as they are often time-consuming, limited in scope, and prone to inconsistencies due to varying levels of taxonomic expertise. Moreover, cryptic species and juvenile life stages are often unidentifiable at lower taxonomic levels. The advent of molecular methods, especially DNA sequencing technologies, has revolutionized biodiversity assessments. The application of DNA sequencing in biodiversity studies began with the Maxam–Gilbert method (Maxam & Gilbert, 1977) and, more prominently, Sanger sequencing (Sanger et al., 1977). Among these, Sanger sequencing became the most widely used because it was technically simpler, produced longer sequences, and was more amenable to automation, laying the groundwork for modern molecular ecology and taxonomic classification (Eren et al., 2022; Shokralla et al., 2014). Although being an essential tool for DNA barcoding of individual specimens, the Sanger sequencing method, however, is unsuitable for characterizing the species communities in a mixture of DNA from environmental samples, except using the tedious cloning step (Agustí et al., 2003; Medinger et al., 2010).

Newer DNA sequencing technologies, known as high-throughput sequencing (HTS), can sequence thousands to millions of DNA fragments simultaneously, making them suitable for use in metabarcoding workflows (Compson et al., 2020; Taberlet et al., 2012). Metabarcoding is a molecular technique that combines DNA barcoding with HTS to identify multiple species from mixed environmental samples (e.g., soil, water, or bulk organism collections) using typically short, standardized genetic markers (Taberlet et al., 2012). In biodiversity assessments, it enables rapid, cost-effective, and non-invasive detection of a broad range of taxa, including elusive or morphologically cryptic species, thereby providing a comprehensive view of community composition (Aylagas et al., 2016; Yu et al., 2012). The advent of practical guidelines for metabarcoding (e.g., Lear et al., 2018; Tedersoo et al., 2022) has enhanced the scalability (i.e., the ability to efficiently process larger numbers of samples or datasets) and throughput of environmental DNA (eDNA) sample processing, thereby increasing the appeal of the metabarcoding method among ecologists. However, second-generation HTS methods, such as those generated by Illumina, Inc. (San Diego, CA, USA), can generate only relatively short reads (a maximum of 2x300 bp in paired-end mode; 2x500 bp since October 2025). Short reads may limit species-level resolution, provide limited phylogenetic depth, and make it challenging to design suitable target-specific primer binding sites (Furneaux et al., 2021; Yang et al., 2020). Standard barcoding regions are generally longer than 500 bp (e.g., the ~650 bp COI gene for animals, ~1.5 kb 16S rRNA gene for bacteria, or 600–1000+ bp ITS region for fungi). In the first wave of metabarcoding studies, Roche's 454 pyrosequencing platform was widely adopted. Although initially limited to read

lengths of 200–300 bp, the technology eventually evolved to support reads of up to 700–1,000 bp, long enough to capture substantial portions of these barcodes. However, 454 sequencing was eventually supplanted by Illumina technology due to its much higher throughput, lower sequencing costs, and superior base calling accuracy (Shokralla et al., 2014). In contrast, Illumina read lengths, although highly accurate, are typically too short to span standard barcode loci in a single fragment, requiring the sequencing of mini-barcodes, which can reduce accuracy and necessitate the use of alternative long-read platforms (Callahan et al., 2019; Liu et al., 2017; Tedersoo et al., 2019). For metabarcoding full-barcodes, there has been a growing interest in third-generation HTS methods (Callahan et al., 2019; Jamy et al., 2020; Karst et al., 2021), such as those offered by Pacific Biosciences (PacBio; Rhoads & Au, 2015) and Oxford Nanopore Technology (ONT; Jain et al., 2016) technologies (Callahan et al., 2019; Jamy et al., 2020; Karst et al., 2021).

The application of these long-read platforms has expanded the scope of metabarcoding toward longer genetic markers, specifically the rRNA operon. Regions of the rRNA operon are widely used markers for fungal and eukaryotic barcoding due to their universal presence, multi-copy nature, and balanced variability. This operon comprises coding regions for the SSU and LSU ribosomal RNAs, interspersed with the more variable, non-coding ITS regions. While the ITS region is the standard barcode for fungi, other regions, such as the SSU and LSU, are also applied, particularly for resolving certain taxonomic groups or when the ITS provides insufficient variation (Heeger et al., 2018; Porras-Alfaro et al., 2014). Sequencing the full-length rRNA (SSU-ITS-LSU) operons using long-read technologies enhances taxonomic resolution by reducing biases inherent to single markers, such as ITS subregions alone (Heeger et al., 2018; J. Lu et al., 2023; Tedersoo, Albertsen, et al., 2021). Nonetheless, sequencing data is frequently susceptible to artifacts and erroneous sequence variants due to PCR and sequencing inaccuracies. For PacBio, errors per individually sequenced molecules are mitigated by forming circular consensus sequences (CCS) – now known as HiFi reads – wherein an amplified genomic locus is circularized and sequenced multiple times to generate highly accurate reads (Hebert et al., 2018). The CCS method effectively optimizes the trade-off between read length and accuracy. It converts raw subreads, which can exceed 100 kb, into HiFi reads, achieving a per-base accuracy comparable to short amplicons' base call accuracy of 99.9% (Castaño et al., 2020). Despite the overall high accuracy, the "raw" data still needs to be run through validation checks via bioinformatics.

The rapid advancement of HTS platforms has been mirrored in the proliferation of software designed to process metabarcoding data (Bolyen et al., 2019). Since a large amount of sequencing data is generated per sample in the metabarcoding workflow, proper bioinformatic processing is necessary to efficiently transform sequences into biodiversity data. A sequence analysis pipeline applies a series of steps using a collection of software and algorithms to transform raw reads into a features table annotated with taxonomic information for downstream analysis. The "features" refer to distinct sequence units that can be

generated using various approaches. For instance, methods that cluster sequences based on a similarity threshold result in features such as operational taxonomic units (OTUs), whereas denoising algorithms resolve amplicon sequence variants (ASVs). Foundational software suites such as mothur (Schloss et al., 2009), USEARCH (Edgar, 2010), and QIIME (Caporaso et al., 2010), include algorithms that can be used to create full metabarcoding data analysis pipelines. Over time, these programs have been augmented with supplementary algorithms to minimize the presence of artifacts and implement various clustering and denoising methodologies. Although the variety of tools provides flexibility, it also presents a challenge, as the number of available bioinformatics pipelines can be overwhelming. Therefore, choosing from the many available pipelines can be confusing. This situation is even more complicated because numerous algorithms in these pipelines were originally designed for short-read sequencing technologies, particularly Illumina, and may not directly translate to the analysis of long-read data. Although metabarcoding may provide high-throughput biodiversity surveys without requiring taxonomic expertise, bioinformatics expertise is still needed to prevent incorrect conclusions from sequence data.

During the bioinformatic processing of sequencing data, initial steps such as quality filtering are applicable to both short- and long-read technologies, as they operate on recorded error probabilities per base. However, dealing with other sequencing artifacts can vary among different data types (depending on the sequencing technologies used). Chimeric DNA fragments, which are artificial sequences resulting from the hybridization of two or more different DNA templates, are among the sequencing artifacts usually present in raw metabarcoding datasets. This phenomenon arises from incomplete template extension or template switching, wherein partially extended sequences hybridize with other templates during PCR amplification (Kebschull & Zador, 2015). Since they are generated in excessive cycles of PCR, they are always less abundant than their parent molecules (Sze & Schloss, 2019) and thus may mostly represent only singleton or doubleton sequences in a sample (Tedersoo et al., 2022). However, not all singletons or doubletons are chimeras. In relatively low-depth long-read datasets (e.g., PacBio), genuine rare taxa may also appear at low frequencies, meaning that indiscriminately discarding these reads risks losing true biological signal. Compared to shorter reads, longer sequences are more at risk of chimera formation, as long amplicons are more likely to be disrupted. They possess more potential chimera breakpoints and may require additional PCR cycles, thereby increasing the risk of chimerism (Heeger et al., 2018; Tedersoo et al., 2018). If these artificial fragments are not removed during the bioinformatic processing of sequence data, the resulting reads can introduce inaccuracies by distorting the diversity estimates in the dataset (Bjørnsgaard Aas et al., 2017; Nilsson et al., 2010). Many metabarcoding pipelines implemented algorithms such as the *removeBimeraDenovo* within DADA2 (Callahan et al., 2016a) and UCHIME (Edgar et al., 2011) for their chimera filtering. While they are effective on short reads, their *de novo* approaches can generate high false-positive rates (Bjørnsgaard Aas et al., 2017; Tedersoo et al., 2022; Tedersoo, Mikryukov, et al., 2021),

yet they are still applied to long-read data (Fichot & Norman, 2013; Furneaux et al., 2021b; Mosher et al., 2013). With the growing adoption of long-read sequencing, it is necessary to evaluate how these algorithms perform beyond their original short-read benchmarks, ensuring that bioinformatic pipelines for long-read metabarcoding yield accurate and reliable biodiversity estimates.

Taxonomic identification of the metabarcoding features is generally the final step in metabarcoding data analysis. This involves comparing the sequences of the features to the reference sequences (database); thus, the accuracy of this step ultimately hinges on the quality of the reference sequence database. Most of the rRNA reference databases widely used were built for short-amplicon workflows and consequently fall short when researchers use long-read sequencing for their eDNA studies. The challenge is compounded by the diversity of molecular markers used across eukaryotes. While studies of prokaryotes have largely standardized the use of the 16S rRNA gene, eukaryotic research employs additional markers, such as the LSU and ITS regions, often depending on the target group of organisms being studied. This has led to the development of specialized, taxonomically curated reference databases that typically focus on a single ribosomal marker and maintain a narrow taxonomic scope. For example, the UNITE database (Abarenkov et al., 2024) is a primary resource for the ITS region for eukaryotes, the PR2 database (Guillou et al., 2013) is focused on SSU data for protists, and the SILVA database (Yilmaz et al., 2014) provides SSU and LSU genes for prokaryotes and eukaryotes. Although these curated databases are great resources, reference databases for full-length SSU-ITS-LSU are lacking. Such full-length SSU–ITS–LSU references would provide consistent marker coverage across diverse eukaryotes, enable cross-validation between regions within the same operon, and minimize taxonomic biases that arise when relying on a single marker.

With the rapid development of bioinformatics pipelines and the growing use of HTS, it is essential to provide researchers with reliable resources and clear instructions. Thus, one of the main goals of this thesis is to thoroughly review the bioinformatics software available for metabarcoding while navigating the main obstacles and choices for amplicon data analysis (I). This thesis also presents EUKARYOME, a new, curated reference database containing full-length rRNA marker sequences, which aims to improve the precision of taxonomic assignment and artifact validation across eukaryotes, thereby meeting the specific requirements of long-read metabarcoding (II). Ultimately, this work critically evaluates the performance of common chimera detection algorithms on long-read amplicons to enhance the integrity of metabarcoding datasets. We also developed a validated workflow to effectively remove artifacts while minimizing the loss of true biological data (III).

This thesis has the following aims:

1) Providing a comprehensive overview of the existing bioinformatic software for metabarcoding analysis and identification of the key challenges and decision points for researchers working with amplicon data. (Paper I)
2) Developing and introducing a curated eukaryotic reference database with full-length rRNA marker sequences to ease high-confidence taxonomic assignment and artifact validation for long-read data. (Paper II)
3) Evaluating the performance of common *de novo* chimera detection algorithms on long DNA marker sequences and developing a validated workflow that minimizes the loss of true biological data while removing artifacts. (Paper III)

2. MATERIALS AND METHODS

2.1. Overview of metabarcoding pipelines

Altogether, thirty-two pipelines were reviewed to provide an overview of available bioinformatics software for the metabarcoding data analysis study (I). The selection criteria were for actively maintained, well-documented software covering various approaches for processing both short- and long-read amplicon sequencing data. Important properties of the metabarcoding software were described, including identifying the software suites and precompiled pipelines, selecting software based on the applied sequencing platform, and the available operating systems and interface preferences. Pipelines that offer a flexible set of individual algorithms were classified as software suites. Others that offer a more automated and predefined chain of analysis steps were classified as precompiled pipelines. The compatibility of each pipeline with various data types was assessed, including single-end and paired-end reads from second-generation platforms such as Illumina, as well as long-read data from third-generation platforms like PacBio. Furthermore, pipelines were evaluated based on their support for various operating systems, including Linux, macOS, and Windows. Furthermore, the types of user interfaces offered were documented, including CLI, GUI, and web-based interfaces, as these factors impact accessibility and ease of use for researchers with varying levels of computational expertise. Important properties of each pipeline were extracted, with a focus on the algorithms employed for essential metabarcoding processing steps. These steps include quality filtering, merging of paired-end reads, error correction (e.g., denoising to generate ASVs), clustering (e.g., to generate OTUs or swarm-clusters), chimera detection and removal, and taxonomic assignment (Figure 1).

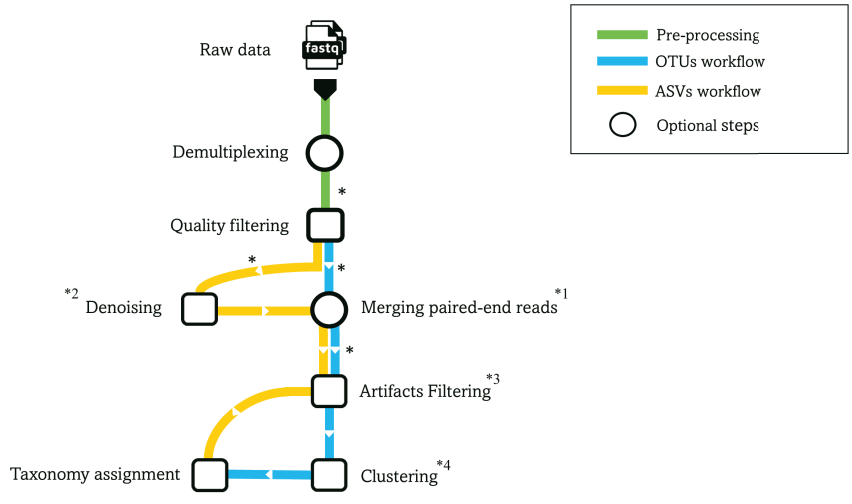

Figure 1. Example of a basic bioinformatics workflow for metabarcoding data. * Primer trimming between any of these steps can be applied. *1 Only for paired-end data (may be performed before or after quality filtering). *2 Error correction; formation of ASVs. *3 Including chimera filtering, off-target gene removal (pseudogene removal, ITS extraction). *4 Formation of OTUs/swarm-clusters. Adapted from the paper I, figure 1.

We also noted whether pipelines were specialized for specific markers, such as 16S rRNA, ITS, or COI. Tool selection is guided by marker specificity, as pipelines are designed either for specific markers or for general-purpose metabarcoding applications.

2.2. EUKARYOME

The construction of the EUKARYOME reference database (II) involved a multistage process of data aggregation from diverse sources, followed by quality control and phylogenetic validation to ensure high accuracy and utility for the community. Initially, a foundational dataset was compiled by aggregating eukaryotic rRNA gene sequences, including the SSU, ITS, and LSU regions, from established and curated databases such as SILVA (version 138.1), PR2 (version 5.1.1), and UNITE (version 10.0). All full-length reads annotated to the species level and published since 2018 were retrieved from INSDC to supplement these reads and capture the most recent taxonomic information. Furthermore, long-read amplicons derived from PacBio and Oxford Nanopore sequence samples from various environments were incorporated (Jamy et al., 2022; M. Hosseyni Moghaddam et al., unpublished results), soil (Tedersoo et al., 2020; M. Sharma et al., unpublished results), marine water (Latz et al., 2022), and animal rumen (Hanafy et al., 2020), as well as data from metagenome and metatranscriptome projects (Galindo et al., 2019; Labarre et al., 2021; Strassert et al., 2019; Tikhonenkov et al., 2022; Torruella et al., 2015). All newly incorporated PacBio sequences were subjected to chimera filtering by UCHIME v4.2 (Edgar et al., 2011) and by using ITSx v.1.1.3 (Bengtsson-Palme et al., 2013) with default options and UNITE, SILVA, and PR2 as a reference. A quality filtering step based on indel distribution was implemented; sequences were aligned in batches using MAFFT v7 (Katoh & Standley, 2013) with standard options, and sequences that exhibited an excessive number of indels within highly conserved rRNA gene regions were flagged as low quality and removed. Due to this aggressive filtering, 15–25% of the initial PacBio reference sequences were eliminated, guaranteeing high fidelity in the final dataset. Oxford nanopore consensus reads from animal, protist, and fungal specimens were analyzed as described earlier, with an exclusion rate of 1%.

For the taxonomic integrity of the EUKARYOME database, the latest taxonomic information was sourced from the curated databases mentioned above. These initial assignments were later improved and revised with BOLD v4 (Ratnasingham & Hebert, 2013) for metazoans, the Outline of Fungi (Wijayawardene et al., 2020), and AlgeaBase (Guiry et al., 2014) for protists. Furthermore, phylogenetic trees, used as a curation strategy, were manually inspected. For the SSU and LSU marker regions, maximum likelihood phylogenies were constructed using IQ-TREE v2.2.2.6 (Minh et al., 2020), which allowed us to visually validate the placement of taxa, identify and discard remaining chimeras, and correct misidentifications frequently inherited from public databases. The database strictly adheres to the main Linnaean ranks (from species to kingdom) to maximize

consistency and cross-study usability. Furthermore, it carefully separates zoological and botanical nomenclature, a critical step to unambiguously resolve the numerous hemihomonyms that confound taxonomic assignments in other major repositories. Dereplication was the final step for the workflow. To form the final, thorough EUKARYOME database, all curated reads were clustered at 100% similarity, and manual curation was performed to remove redundant sequences, retaining a single high-quality representative sequence for each species. For making a database easily convertible for any taxonomy assignment tool, a pipeline was developed in Python and Snakemake (Köster et al., 2021) workflow manager.

2.3. Chimera filtering algorithms evaluation

The performance of *de novo* chimera detection algorithms *uchime_denovo*, *chimeras_denovo* (both, as implemented in VSEARCH v2.29.4), and *removeBimeraDenovo* (DADA2 v.1.32) on long-read amplicons was evaluated (III) using a dual-pronged approach combining a simulated dataset and an empirical full-length rRNA ITS dataset from diverse environmental sources. To test the precision and recall of the algorithms, a simulated PacBio dataset was first constructed using SimLoRD v1.0.4 based on 186 full-length ITS reference sequences obtained from the EUKARYOME v.1.9.2 database (II). This generated 424,010 reads with characteristic PacBio error profiles, which were then quality filtered by DADA2 v.1.32 and processed with ITSx to yield a clean dataset of 44,470 full-length ITS sequences. A custom Python script (https://github.com/alihkz94/long-chimeric-reads-project/blob/main/Simulated_data/chimera_generator.py) was subsequently used to generate 2,484 *in silico* chimeras into this dataset, creating a final test set with a known 6.2% chimera rate. We tested the precision and recall of *uchime_denovo* and *chimeras_denovo* by tuning with different parameter settings (Table 1) in each run with our simulated dataset. The effectiveness of each algorithm was quantified using the F1 score, a harmonic mean of precision and recall, calculated as:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision measures the proportion of true chimeras among all identified chimeras, and recall measures the proportion of true chimeras that were correctly identified. We conducted 49 runs using the *uchime_denovo* and 22 runs using the *chimeras_denovo*.

Table1. List of modified parameters in *uchime_denovo* and *chimeras_denovo* to examine their precision and recall in the simulated dataset.

	Parameter	Default Value	Tested ranges	Description
uchime_denovo	--dn	1.4	1.4–2.0 (step 0.2)	Pseudo-count cutoff for nucleotide difference before assigning as chimeric. A higher value increases specificity, reducing false positives.
	--mindiffs	3	2–4 (step 1)	Minimum differences required per query segment to consider a chimera. Higher values reduce the likelihood of detecting chimeras.
	--minh	0.28	0.10–0.28 (step 0.02), 0.10–0.05 (step 0.01)	Threshold for considering a sequence as chimera. Smaller values improve sensitivity.
	--abskew	2	2–16 (step 1)	Minimum abundance ratio between the parent sequence and chimera. Reflects abundance skew in the dataset.
	--xn	2	2, 3	Controls how strongly the algorithm weighs instances where a sequence is not classified as a chimera.
	--mindiv	0.8	0.4, 0.6	The divergence threshold below which sequences are not considered chimeric. Lower values increase sensitivity.
chimeras_denovo	--chimeras_diff_pct	0	0.5–0.9	Percentage mismatch allowed in chimeric regions. Larger values increase sensitivity to sequence mismatches.
	--chimeras_ length_min	10	10–60 (step 10)	Minimum length of chimeric regions. Longer regions reduce false positives.
	--chimeras_parts	1	2, 3	The number of parts a sequence is divided into and adjusted to test segmentation effects on detection.
	--abskew	1	2–6 (step 1)	Minimum abundance ratio

For the empirical dataset, a dataset from Jamy et al. (2022, BioProject PRJEB45931 in ENA) was utilized, comprising 10,609,939 PacBio Sequel II reads of full-length ITS amplicons from 18 environmental samples, including marine, freshwater, and soil habitats. The raw data underwent a pre-processing workflow before chimera filtering analysis, which included primer trimming with cutadapt v4.4 (Martin, 2011), quality filtering using DADA2 v.1.32 (truncLen = 0, maxEE = 2, minQ = 3), ITS region extraction via ITSx (--nhmmer TRUE, -E 1e-2, --complement TRUE, --only_full TRUE), dereplication with VSEARCH, and filtering of putative tag-jumps using UNCROSS2 (Edgar, 2018) within PipeCraft2 (Anslan et al., 2017) (options: f = 0.03, p = 1) (Figure 2). This yielded 2,070,676 dereplicated ITS sequences that were used for comparative analysis of the chimera filtering algorithms. Chimera filtering outputs (chimeric & non-chimeric) were subjected to BLASTn (Altschul et al., 1997) search (word size = 7; reward = 1; penalty = −1; gap opening cost = 1; gap extension cost = 2), using EUKARYOME v.1.9.2 as a reference database. A putative chimera was classified as a false-positive chimera if it showed high identity (≥99%) and high query coverage (≥99%) to a reference sequence. Conversely, the reads in the non-chimeric output were identified as false-negative chimeras if their best BLASTn hit consisted of multiple alignments with the first HSP covering less than 85% of the query.

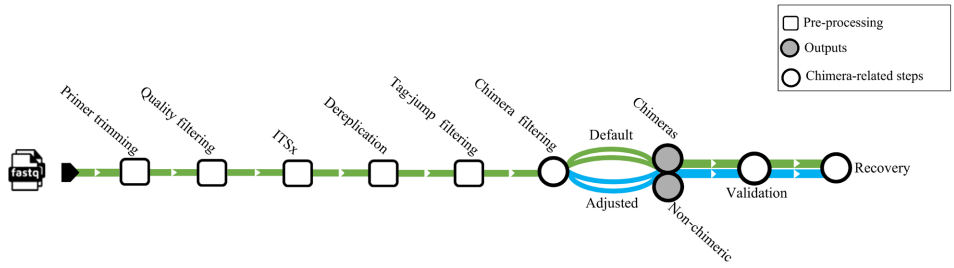

Figure 2. Bioinformatic analysis workflow related to the empirical dataset. Adapted from paper III, figure 1.

To evaluate the impact of different chimera filtering strategies on the related dataset, OTUs were clustered at a 98% similarity threshold using VSEARCH. Community composition was compared across multiple datasets: raw (without chimera filtering), default filtering, and adjusted settings with correction for false positives and false negatives ("adjusted + FP – FN"). The raw OTU count tables were first transformed using a presence/absence (PA) via the *decostand* function in the R package vegan v.2.6-8 (Oksanen et al., 2024). Bray-Curtis dissimilarity matrices were then computed using the *vegdist* function. To visualize community differences, NMDS ordinations were created using these matrices. Pairwise Procrustes tests were applied to the NMDS ordinations to quantify the ordination-space concordance between various filtering strategies. A linear mixed-effects model was fitted using the lme4 package in R to further examine the effects of parameter tuning on OTU richness. This model evaluated the effects of the

sample's isolation source (forest soil, lake water, marine water, and peat soil) as fixed factors and the chimera filtering method (*chimeras_denovo*, *removeBimeraDenovo*, *uchime_denovo*) on Procrustes residuals (the response variable, representing the discrepancy between community ordinations). The sample was included as a random effect since there were biological replicates. A likelihood-ratio test against a reduced model was used to confirm the significance of the interaction term between method and source. Finally, the effect of different chimera removal algorithms on alpha diversity was evaluated. Kruskal-Wallis rank-sum test was performed on Shannon diversity index. We fitted a linear model with the chimera removal method, isolation source, and their interaction as fixed effects to account for potential confounding variables. The obtained residuals were then used as an adjusted measure of alpha diversity. Their structural traits were analyzed to determine why some chimeric sequences were missing from the filtration (i.e., false negatives). We used the Infernal software (v.1.1.5) (Nawrocki & Eddy, 2013) along with the inferrnal R package (v.0.99.8) (Furneaux, 2025). We determined the position and the number of the 5.8S rRNA gene in related sequences. Furthermore, a genuine ITS region contains only one 5.8S gene, so finding multiple copies within a sequence suggests a chimeric artifact formed from joined ITS regions. Thus, the length distribution of these false negatives was analyzed, as unusually long sequences may indicate concatenation events. All statistical analyses were performed in R v.4.4.2 (R Core Team, 2024), and visualizations were produced with ggplot2 v.3.5.1 (Wickham, 2016).

3. RESULTS AND DISCUSSION

The rapid development of HTS has transformed biodiversity assessments and advanced the field beyond the constraints of traditional methods. The generation of massive datasets by these machines has led to the proliferation of bioinformatic software, which can be challenging to navigate. Though many of these tools were developed for short-read data, the field is increasingly adopting long-read sequencing (e.g., by using full barcodes) to achieve greater taxonomic resolution. Therefore, long-read technologies offer improved taxonomic and phylogenetic resolution in metabarcoding by enabling full-length barcode sequencing without the biases introduced by fragmentation, although they are more expensive per sequenced base pair and have platform-specific error profiles. Short-read platforms are still useful for high-throughput and cost-sensitive studies that prioritize per-base accuracy. This shift, however, introduces new analytical hurdles, including an increased risk of chimeric artifacts and a scarcity of curated, full-length reference databases required for accurate taxonomic assignments. Therefore, a series of studies (I, II, III) was conducted to address these challenges. This work first provides a comprehensive overview of the available bioinformatics software (I), then introduces EUKARYOME, a reference database designed to meet the specific requirements of long-read technologies (II), and finally evaluates common chimera detection algorithms to develop a validated workflow that improves the integrity of metabarcoding datasets (III).

3.1. Metabarcoding bioinformatics pipelines

The overview of metabarcoding pipelines (I) included 32 distinct software programs (Table 2), broadly divided into two categories based on their flexibility: software suites (Bolyen et al., 2019; Boyer et al., 2016; Callahan et al., 2016; Rognes et al., 2016) and precompiled pipelines (Abdala Asbun et al., 2020; Albanese et al., 2015; Curd et al., 2019; Palmer et al., 2018). APSCALE (Buchner et al., 2022), gDAT (Vasar et al., 2021), CoMa (Hupfauf et al., 2020), PipeCraft2, SEED2 (Větrovský et al., 2018) pipelines provide predefined analytical steps (Figure 1) with tunable parameter settings, thus facilitating analyses for researchers with limited bioinformatics expertise. Among software suites, VSEARCH was particularly prominent due to its comprehensive functionality, which often led to its incorporation into predefined pipelines, making it particularly appealing when dealing with large datasets. Both mothur and QIIME 2 (Bolyen et al., 2019) have incorporated various functionalities by leveraging algorithms from tools like VSEARCH and DADA2, achieving a balance between versatility and user-friendliness.

A standard metabarcoding workflow (Figure 1) includes several main steps. It begins with a demultiplexing step (integrated into 19 pipelines) by using tools like cutadapt or sdm (Hildebrand et al., 2014) and primer trimming (with

cutadapt, Trimmomatic (Bolger et al., 2014), or AdapterRemoval (Lindgreen, 2012)). This is followed by quality filtering and the merging of paired-end reads (only in Illumina and DNBSeq), which can range from sliding-window methods to advanced error modeling, such as that used in DADA2. The workflow then proceeds with artifact filtering, where chimeras are removed using *de novo* or reference-based methods; some pipelines, such as FROGS (Bernard et al., 2021) and NextITS (Mikryukov et al., 2025) even include modules to recover sequences that may have been incorrectly detected as chimeras (false-positive chimeras). The final stages are feature generation and taxonomy assignment. Feature generation creates units such as ASVs (DADA2 or UNOISE (Edgar, 2016)), OTUs (e.g., VSEARCH, CD-HIT (Li & Godzik, 2006), OTUCLUST (Albanese et al., 2015), CROP (Hao et al., 2011), etc.) or swarm-clusters (Mahé et al., 2021). Eventually, the taxonomy assignment classifies these features using alignment-based (Altschul et al., 1997) or composition-based classifiers (Wang et al., 2007).

The data structure, sequencing platform, and analyzed marker are among the most relevant factors influencing the selection of applied bioinformatics, thus a metabarcoding pipeline. Most pipelines were originally designed for Illumina paired-end sequencing, where merging forward and reverse reads improves error correction and chimera detection, as in mothur, QIIME 2, and DADA2. Regarding marker-specific pipelines, the analysis of 16S amplicons is broadly supported across multiple pipelines. In contrast, since the ITS region is highly variable in length among eukaryotic groups, it complicates the bioinformatics analysis steps that rely on aligning (such as, e.g., mothur OTU clustering) or require uniform sequence length (such as, e.g., deblur). Pipelines such as PIPITS (Gweon et al., 2015) and DAnIEL (Loos et al., 2021) are specifically designed for ITS1/2 amplicon analyses; however, they are only process paired-end Illumina data. Processing COI amplicons presents different challenges, as nuclear mitochondrial pseudogenes (NUMTs) must be filtered out; solutions for this are implemented in MetaWorks (Porter & Hajibabaei, 2022), PipeCaft2, and VTAM (González et al., 2023). More recently, the HAPP pipeline (Sundh et al., 2024) introduced a new algorithm, NEEAT, which combines echo detection, evolutionary signal analysis, and read abundance thresholds to distinguish real sequences from NUMTs. Apart from the full pipelines, the multisample features matrix may be processed with metaMATE (Andújar et al., 2021) to remove putative NUMTs and other erroneous sequences (based on, e.g., length and relative read abundance). For the analysis of PacBio long-read data, which offers full-length barcodes such as the 16S rRNA or ITS regions, several pipelines have been adapted to the unique characteristics of HiFi reads. For instance, the NextITS pipeline, designed for full-length fungal ITS, implemented the correction of homopolymer errors and UNCROSS2 to reduce cross-sample contamination caused by tag-jumping events. DADA2 is integrated into several pipelines recommended for long-read amplicon analysis, including QIIME2, dadasnake, nf-core/ampliseq, and PipeCraft2. Notably, PipeCraft2 incorporates the NextITS pipeline alongside DADA2, which combines both general-purpose and ITS-specific denoising in a unified framework optimized for long-read sequences.The DADA2 pipeline (Callahan et al.,

2019) has been adapted to incorporate an error model specifically designed for PacBio reads, enabling accurate denoising by modeling platform-specific error profiles. Although DADA2 has a specific denoising algorithm that performs well on synthetic long reads (Callahan et al., 2021), its application may still require higher sequencing depth or relaxed denoising parameters for high diversity samples (Furneaux et al., 2021), especially in complex communities (e.g., soil). Importantly, other denoising tools such as UNOISE or Deblur (Amir et al., 2017) are not optimized for third-generation sequencing machines.

The formation of features in many pipelines includes both ASVs and OTUs. The choice between ASVs and OTUs is context-dependent, with potential for combined use in studies, highlighting key trade-offs in resolution, reproducibility, and error handling. ASVs provide fine-scale, biologically informative resolution that is lost in OTU clustering (Callahan et al., 2016). Denoisers inherently discard low-abundant variants as artifacts (Anslan et al., 2021; Reitmeier et al., 2021), thereby reducing spurious features (e.g., De Santiago et al., 2022); however, this approach risks the loss of rare taxa in low-depth datasets. Sensitivity to rare ASVs can be adjusted in pipelines like DADA2, FROGS, VSEARCH, and USEARCH. ASVs ensure stable, reproducible units across studies, unlike dataset-specific OTUs (Callahan et al., 2017). However, ASVs may misrepresent species in taxa with high intraspecific polymorphism, such as metazoan COI (Brandt et al., 2021) or fungal ITS multicopy/size variation (Estensmo et al., 2021; Tedersoo et al., 2022), unless specialized handling is applied (e.g., in FROGS). Additionally, post-ASV clustering addresses this (Antich et al., 2021; Brandt et al., 2021; Porter & Hajibabaei, 2020), as implemented in MetaWorks, PipeCraft2, and dadasnake (Weißbecker et al., 2020). Additionally, QIIME 2, nf-core/ampliseq (Ewels et al., 2020; Straub et al., 2020), and LotuS2 enable feature collapse by annotated taxon levels, producing taxa features. Overall, community patterns remain similar across feature types (e.g., Glassman & Martiny, 2018; Kang et al., 2021; Porter & Hajibabaei, 2020), though rare taxa recovery varies (Nearing et al., 2018). Post-clustering tools such as LULU (Frøslev et al., 2017) could be useful for reducing inflated richness estimates by identifying and merging erroneous "daughter" sequences that co-occur with a more abundant "parent" sequence. Post-clustering tools, such as LULU, are implemented in PipeCraft2, AMPTk (Palmer et al., 2018), eDNAflow (Mousavi-Derazmahalleh et al., 2021), LotuS2, and ReClustOR (Terrat et al., 2020), in BIOCOM-PIPE (Djemiel et al., 2020). Furthermore, VTAM (González et al., 2023) features a control-based workflow for distinguishing rare true taxa from artifacts in low-depth, long-read datasets. Overall, while community patterns are often highly similar regardless of feature type, the choice between ASVs and OTUs should reflect study-specific priorities regarding taxonomic resolution, sequencing depth, and marker gene characteristics. Regardless of the selected approach, implementing post-clustering curation and appropriate filtering strategies remains essential for balancing the detection of rare taxa against the removal of artifactual sequences.

Table 2. A list of reviewed metabarcoding data analysis pipelines, including their features, demultiplexing steps, primer removal integrity, and the designed marker types.

Pipelines	Feature	Demux	Primer/ adapter removal	Marker
AMPtk	ASV, OTU	yes	yes	16S, 28S, COI, ITS
Anacapa	ASV	multilocus demux based on primers	yes	Multi-marker
APSCALE	ASV, OTU	no	yes	Multi-marker
Barque	ASV, OTU	no	yes	Multi-marker
BIOCOM-PIPE	ASV, OTU	yes	yes	16S, 18S, 23S
Cascabel	ASV, OTU, swarm-cluster	yes	yes	Multi-marker
CoMA	ASV, OTU, swarm-cluster	no	yes	Multi-marker
DADA2	ASV	no	yes	Multi-marker
Dadaist2	ASV	no	yes	Multi-marker
dadasnake	ASV, OTU	no	yes	Multi-marker (with ITSx)
DAnIEL	ASV	yes	yes	ITS
eDNAflow	ASV	yes	yes	Multi-marker
FROGS	ASV, swarm-cluster	yes	yes	Multi-marker (with ITSx)
gDAT	OTU	no	yes	18S, ITS
JAMP	OTU	yes	yes	Multi-marker (without taxonomic assignment)
LotuS2	ASV, OTU, swarm-cluster	yes	yes	Multi-marker (16S, 18S, 23S, 28S, ITS with ITSx)
MetaWorks	ASV, OTU	no	yes	Multi-marker; SSU (12S, 16S, 18S), ITS (with ITSx), LSU (28S), COI (with pseudogene removal), rbcL (with pseudogene removal)
MICCA	ASV, OTU, swarm-cluster	yes	yes	16S, 18S, 28S, ITS

Pipelines	Feature	Demux	Primer/ adapter removal	Marker
mothur	ASV, OTU	yes	no	Multi-marker
NextITS	ASV, OTU, swarm-cluster	yes	yes	ITS (with ITSx)
nf-core/ ampliseq	ASV	no	yes	Multi-marker (with ITSx)
OBITools3	ASV	yes	yes	Multi-marker
PEMA	OTU, swarm-cluster	no	yes	Multi-marker
PipeCraft2	ASV, OTU	yes	yes	Multi-marker (with ITSx and pseudogene removal)
PIPITS	OTU	no	yes	ITS (with ITSx)
QIIME 2	ASV, OTU	yes	yes	Multi-marker (with ITSxpress)
SCATA	OTU	yes	yes	Multi-marker (with ITSx)
SEED2	OTU	yes	yes	16S, ITS (with ITSx)
Tourmaline	ASV, OTU	yes	yes	Multi-marker
USEARCH	ASV, OTU	no	no	Multi-marker
VSEARCH	ASV, OTU	no	no	Multi-marker
VTAM	ASV, OTU	yes	yes	Multi-marker (with pseudogene removal)

3.2. EUKARYOME

EUKARYOME (available at http://www.eukaryome.org) is a curated database for rRNA markers of all eukaryotes (II), developed and maintained by the Mycology and Microbiology Center (MMC) at the University of Tartu. In its last major release (v2.0), EUKARYOME contains nearly 1.3 million entries, and it features dedicated datasets for the SSU (353,679 sequences), ITS (1,597,427 sequences), and LSU (377,679 sequences) markers. The key innovation is that it includes 193,705 full SSU-ITS-LSU operon sequences. This aspect is crucial for reference-based chimera filtering and taxonomic assignment of long amplicons. To maximize its utility and facilitate widespread adoption, the resource provides pre-formatted datasets for direct use with prevalent bioinformatics pipelines/ tools, such as QIIME 2, mothur, SINTAX, DADA2, and UCHIME. For latter formatting, I developed the SnakeEUK pipeline (https://github.com/alihkz94/SnakeEUK). Read variations among subset databases related to each marker (e.g., SSU, LSU,

ITS) reflect taxonomic and metabarcoding efforts across different groups. For instance, ITS dominates the dataset since it reflects its common use in fungal barcoding, while SSU and LSU numbers depict targeted efforts in metazoans, protists, and arbuscular mycorrhizal fungi. During the specific curation process integrated for EUKARYOME, approximately 4,256 (0.4%) of tested reads from the INSDc database were detected as chimeric or low quality. Consequently, EUKARYOME offers a broader taxonomic scope by incorporating recent data from metagenomics, metatranscriptomics, and long-read metabarcoding studies, encompassing over 172,000 species across 36 eukaryotic kingdoms, thereby significantly enhancing its coverage compared to other commonly used databases. As an example, the Glomeromycota phylum subset was analyzed. For the SSU marker, which roughly half of all Glomeromycota metabarcoding studies use, EUKARYOME increases the identifiable genera from 18 to 37 for SSU, 20–21 to 38 for ITS, and 21–35 to 42 for LSU, complementing existing databases such as MaarjAM (Öpik et al., 2010) and GlobalAMFungi (Větrovský et al., 2023). Furthermore, the curation process identified areas for potential improvement in publicly available datasets, pinpointing sequences within INSDc and MaarjAM that could benefit from further quality filtering, including 21.6% in the SSU, 17.2% in the ITS, and 16.9% in the LSU datasets.

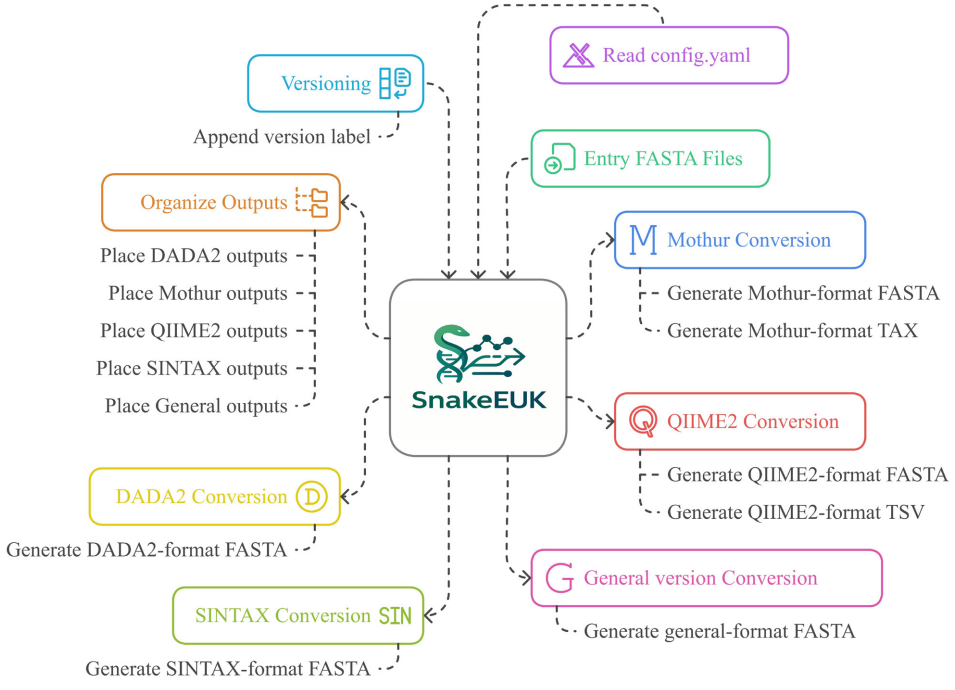

Figure 3. The SnakeEUK pipeline workflow. Pipeline processes the FASTA file to perform format-specific conversions (mothur, QIIME2, DADA2, SINTAX, general) and organizes outputs with versioning.

3.3. Evaluation of chimera filtering algorithms on long-read amplicons

To further refine the bioinformatic workflow of long-read amplicons, three commonly used *de novo* chimera detection algorithms, *uchime_denovo, chimeras_denovo,* and *removeBimeraDenovo* were evaluated on simulated and empirical long-read ITS datasets to determine their precision, recall, and their effects on final OTU composition and community structure (III). On a simulated + empirical dataset, we developed a validating and recovering workflow to minimize data loss through discarding false positive chimeras.

In the simulated dataset, the default settings of *uchime_denovo* yielded a comparatively high precision (F1 = 0.87), detecting a significant portion of the simulated chimeric reads. Adjusting the algorithm's settings yielded only a modest increase in precision (F1 = 0.89). In contrast, the *chimeras_denovo* default settings identified true chimeras but flagged a large number of sequences as false positives, indicating overly aggressive filtering that misclassified real sequence variation as chimeric. Optimizing the chimera filtering settings significantly improved its performance. However, despite these adjustments, false-positive (FP) detections remained significantly higher than in *uchime_denovo*. The *removeBimeraDenovo*, only its default settings were applied; the performance was similar to *chimeras_denovo*'s default settings, detecting numerous false positives alongside true chimeras. Undetected chimeras (false negatives) across *uchime_denovo* and *chimeras_denovo* were predominantly abnormally long (>1000 bp) or short (<100 bp), indicating challenges in capturing extreme length artifacts during de novo detection.

In the empirical dataset, the performance of chimera detection varied substantially across methods and parameter settings, highlighting the complexity of optimizing these algorithms for real-world applications. Only 151 reads were commonly flagged as chimeric by all three methods under default conditions, revealing that algorithms not only differ in their recall (sensitivity) level, but also fundamentally disagree on which sequences are chimeras in the first place. However, parameter optimization proved challenging as adjustments that improved performance on simulated data did not translate directly to the same efficiency on empirical data. As Edgar et al. (2011) demonstrated, the UCHIME's filtering efficiency can decline in highly complex environmental samples. False-positive rates revealed significant method-dependent biases. For instance, *uchime_denovo* detected far fewer false positives with default settings, whereas *chimeras_denovo* and *removeBimeraDenovo* detected much higher rates. Adjusted settings improved *chimeras_denovo* performance by reducing false positives, but increased *uchime_denovo* false positives, illustrating the trade-off between precision and recall. *uchime_denovo* consistently misidentified certain fungal taxa as chimeras across all habitats. In contrast, the *chimeras_denovo* method exhibited habitat-specific biases toward non-fungal taxa. The *removeBimeraDenovo* showed similar biases to *uchime_denovo* for fungal taxa in soil, as well as for Annelida in marine

samples. False-negative (FN) rates remained consistently low across methods and under both parameter settings, but the limited overlap of identified chimeric sequences among algorithms underscored the method-specific nature of chimera detection. Similar results were reported in a recent benchmarking study (Overgaard et al., 2024) showing that filtering PacBio HiFi reads with adjusted *uchime_denovo* parameters can achieve high accuracy, though our findings emphasize that such optimization is highly context-dependent and dataset-specific. Moreover, sequence overlap revealed that all algorithms missed 1,240 chimeric sequences in the empirical dataset under default settings, and 1,238 under adjusted settings. These consistently missed sequences exhibited a bimodal length distribution with peaks at approximately 500 bp and 5000 bp. Longer sequences (median ~4500 bp) contained multiple 5.8S rRNA gene regions because these chimeras formed during PacBio library preparation, rather than during PCR amplification. Observations suggest that long chimeric constructs that emerge during SMRTbell adaptor ligation can evade standard detection algorithms (Fichot & Norman, 2013; Griffith et al., 2018). Comparing default and adjusted + FP – FN settings within algorithm pairs revealed a high correlation, with *uchime_denovo* displaying the strongest concordance. Linear mixed-effects modeling confirmed the significant effects of both chimera removal methods and isolation source (forest soil, lake water, marine water, and peat soil), with a highly significant interaction between method and source. Furthermore, *uchime_denovo* demonstrated consistent performance across environments, while *chimeras_denovo* showed the largest changes in the final community (OTUs) structure, especially in forest soil and lake water samples. Similar to our findings, method-specific taxonomic biases were observed in the false-positive analysis.

Overall, the high false-positive rate in the tested algorithms had a minor impact on community-level analysis. We found no significant differences in the Shannon diversity index and community composition between datasets where false positives and false negatives were corrected and where the latter were not corrected. However, secondary validation integration can be useful for improving long-read data analysis by rescuing false-positive reads and refining taxonomic resolution at the species level (Christel et al., 2023; Hu et al., 2022; N. Lu et al., 2023). Furthermore, the pairwise Procrustes comparisons revealed that applying the adjusted + FP – FN correction minimized method-specific biases, resulting in stronger correlations between them. Therefore, I implemented a validation module in PipeCraft2 that applies a reference-based rescue step. This integration eliminates the need for manual corrections outside of the workflow. Our results suggest that chimera filtering complexity increases with environmental sample diversity. This makes secondary validation particularly important for high-complexity substrates, such as forest soils, compared to marine waters. With growing recognition that error-free chimera filtering is unachievable (Edgar, 2016), integrating diverse secondary validation strategies into metabarcoding pipelines represents a balanced approach to managing the inherent trade-offs between precision and recall in chimera detection.

3.4. Future perspectives

The review of bioinformatics pipelines conducted in this thesis (I) demonstrates the current lack of a universal standard for metabarcoding data analysis. With the increasing adoption of third-generation sequencing, it is evident that the selection of an optimal workflow is contingent upon the specific characteristics of the dataset, the target taxonomic groups, and the sequencing chemistry used. Based on the evaluation of software accessibility (I), the development of reference databases (II), and the characterization of long-read artifacts (III), the field appears to be transitioning toward more modular and adaptable analytical approaches. Such frameworks emphasize flexibility in tool integration, reproducibility through workflow managers, and optimization for long-read platforms, enabling researchers to tailor analyses while minimizing bottlenecks in processing complex datasets. After reviewing a wide range of existing pipelines, no single "ideal" workflow has emerged that consistently outperforms others in all scenarios. Instead, the most effective approaches currently combine strong components such as high-accuracy denoising (e.g., DADA2 adapted for long reads), reliable chimera filtering with secondary validation, and comprehensive taxonomic assignment using curated multi-marker databases like EUKARYOME within modular, reproducible frameworks managed by tools such as Snakemake or Nextflow. Nevertheless, these combinations reflect a set of emerging best practices that have gained broad acceptance through comparative evaluations and community guidelines, including rigorous artifact filtering, hybrid ASV/OTU strategies for variable markers, incorporation of mock communities or controls for validation, and emphasis on reproducibility. We are still distant from a "unified standardized pipeline", as ongoing challenges include platform-specific error profiles (sequencing chemistry and platforms are in continuous development), incomplete reference coverage for rare taxa, and the need for marker-agnostic standardization. Nevertheless, ongoing developments in user-friendly, open-source pipelines suggest that a more unified and integrated framework could emerge within the next 5–10 years, particularly as long-read accuracies continue to improve and community efforts converge on best practices for validation, benchmarking, and interoperability. Achieving this would require coordinated standardization initiatives, expanded high-quality reference resources, and broader adoption of workflow managers to facilitate sharing and iteration.

4. CONCLUSIONS

The following main conclusions can be inferred from my thesis:

- A comprehensive overview of 32 bioinformatics pipelines revealed significant diversity in workflow structures where the choice of a tool depends on the sequencing platform, underlying data structure, and marker. This overview provides a practical guide for selecting appropriate metabarcoding analysis tools (I).
- The EUKARYOME database was introduced as the first curated reference database for all eukaryotes containing full-length rRNA operon sequences, improving taxonomic identification accuracy and chimera validation for long-read data (II).
- Evaluation of the commonly used *de novo* chimera detection algorithms on metabarcoding data reveals high false-positive rates when applied to long reads. Whereas the overall community patterns are not severely affected, the higher precision requires parameter tuning and a secondary validation workflow to effectively remove artifacts while minimizing the loss of true biological data. Parameter tuning and secondary validation workflows were shown to minimize the loss of genuine biological sequences while removing artifacts. However, the impact on broad-scale community structure remained limited, indicating that method-specific biases are more pronounced in taxonomic composition than in overall diversity patterns (III).

REFERENCES

Abarenkov, K., Nilsson, R. H., Larsson, K. H., Taylor, A. F. S., May, T. W., Frøslev, T. G., Pawlowska, J., Lindahl, B., Põldmaa, K., Truong, C., Vu, D., Hosoya, T., Niskanen, T., Piirmann, T., Ivanov, F., Zirk, A., Peterson, M., Cheeke, T. E., Ishigami, Y., … Kõljalg, U. (2024). The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Research*, *52*(D1), D791–D797. https://doi.org/10.1093/NAR/GKAD1039

Abdala Asbun, A., Besseling, M. A., Balzano, S., van Bleijswijk, J. D. L., Witte, H. J., Villanueva, L., & Engelmann, J. C. (2020). Cascabel: A Scalable and Versatile Amplicon Sequence Data Analysis Pipeline Delivering Reproducible and Documented Results. *Frontiers in Genetics*, *11*, 489357. https://doi.org/10.3389/FGENE.2020.489357/BIBTEX

Agustí, N., Shayler, S. P., Harwood, J. D., Vaughan, I. P., Sunderland, K. D., & Symondson, W. O. C. (2003). Collembola as alternative prey sustaining spiders in arable ecosystems: Prey detection within predators using molecular markers. *Molecular Ecology*, *12*(12), 3467–3475. https://doi.org/10.1046/J.1365-294X.2003.02014.X;WGROUP:STRING:PUBLICATION

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., & Donati, C. (2015). MICCA: A complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports*, *5*. https://doi.org/10.1038/srep09743

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/NAR/25.17.3389

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., & Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources*, *21*(6), 1772–1787. https://doi.org/10.1111/1755-0998.13337

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–e240. https://doi.org/10.1111/1755-0998.12692

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science*, *3*(JUN), 204869. https://doi.org/10.3389/FMARS.2016.00096/BIBTEX

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., … Nilsson, R. H. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution*, *4*(10), 914–919. https://doi.org/10.1111/2041-210X.12073

Bernard, M., RuCrossed D sign©, O., Mariadassou, M., & Pascal, Gc. D. sign©raldine. (2021). FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. *Briefings in Bioinformatics*, *22*(6), 1–6. https://doi.org/10.1093/BIB/BBAB318

Bjørnsgaard Aas, A., Davey, M. L., & Kauserud, H. (2017). ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources*, *17*(4), 730–741. https://doi.org/10.1111/1755-0998.12622

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. In *Nature Biotechnology* (Vol. 37, Issue 8, pp. 852–857). Nature Publishing Group. https://doi.org/10.1038/s41587-019-0209-9

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Buchner, D., Macher, T. H., & Leese, F. (2022). APSCALE: advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. *Bioinformatics*, *38*(20), 4817–4819. https://doi.org/10.1093/BIOINFORMATICS/BTAC588

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

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research*, *47*(18), E103. https://doi.org/10.1093/NAR/GKZ569

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pẽa, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. In *Nature Methods* (Vol. 7, Issue 5, pp. 335–336). https://doi.org/10.1038/nmeth.f.303

Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B. D., Stenlid, J., Clemmensen, K. E., & Olson, Å. (2020). Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities. *New Phytologist*, *228*(3). https://doi.org/10.1111/nph.16731

Christel, A., Dequiedt, S., Chemidlin-Prevost-Bouré, N., Mercier, F., Tripied, J., Comment, G., Djemiel, C., Bargeot, L., Matagne, E., Fougeron, A., Mina Passi, J. B., Ranjard, L., & Maron, P. A. (2023). Urban land uses shape soil microbial abundance and diversity. *Science of The Total Environment*, *883*, 163455. https://doi.org/10.1016/J.SCITOTENV.2023.163455

Compson, Z. G., McClenaghan, B., Singer, G. A. C., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale. *Frontiers in Ecology and Evolution*, *8*. https://doi.org/10.3389/fevo.2020.581835

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., & Meyer, R. S. (2019). Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

Edgar, R. C. (n.d.). *UCHIME2: improved chimera prediction for amplicon sequencing*. https://doi.org/10.1101/074252

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

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257. https://doi.org/10.1101/081257

Edgar, R. C. (2018). UNCROSS2: identification of cross-talk in 16S rRNA OTU tables. *BioRxiv*, 400762. https://doi.org/10.1101/400762

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/BIOINFORMATICS/BTR381

Eren, K., Taktakoğlu, N., & Pirim, I. (2022). DNA Sequencing Methods: From Past to Present. *Eurasian Journal of Medicine*, *54*, S47–S56. https://doi.org/10.5152/EURASIANJMED.2022.22280

Fichot, E. B., & Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*, *1*(1), 1–5. https://doi.org/10.1186/2049-2618-1-10/FIGURES/4

Furneaux, B. (2025). *infernal: Interface to Call Programs from Infernal RNA Covariance Model Package*. https://github.com/brendanf/infernal

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021a). Long- and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities. *Molecular Ecology Resources*, *21*(6), 1833–1849. https://doi.org/10.1111/1755-0998.13387

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021b). Long- and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities. *Molecular Ecology Resources*, *21*(6), 1833–1849. https://doi.org/10.1111/1755-0998.13387

Galindo, L. J., Torruella, G., Moreira, D., Eglit, Y., Simpson, A. G. B., Völcker, E., Clauß, S., & López-García, P. (2019). Combined cultivation and single-cell approaches to the phylogenomics of nucleariid amoebae, close relatives of fungi. *Philosophical Transactions of the Royal Society B*, *374*(1786). https://doi.org/10.1098/RSTB.2019.0094

González, A., Dubut, V., Corse, E., Mekdad, R., Dechatre, T., Castet, U., Hebert, R., & Meglécz, E. (2023). VTAM: A robust pipeline for validating metabarcoding data using controls. *Computational and Structural Biotechnology Journal*, *21*, 1151–1156. https://doi.org/10.1016/J.CSBJ.2023.01.034

Griffith, P., Raley, C., Sun, D., Zhao, Y., Sun, Z., Mehta, M., Tran, B., & Wu, X. (2018). PacBio library preparation using blunt-end adapter ligation produces significant artefactual fusion DNA sequences. *BioRxiv*, 245241. https://doi.org/10.1101/245241

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., De Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/NAR/GKS1160

Guiry, M. D., Guiry, G. M., Morrison, L., Rindi, F., Miranda, S. V., Mathieson, A. C., Parker, B. C., Langangen, A., John, D. M., Bárbara, I., Carter, C. F., Kuipers, P., & Garbary, D. J. (2014). AlgaeBase: An On-line Resource for Algae. https://Doi.Org/10.7872/Crya.V35.Iss2.2014.105, *35*(2), 105–115. https://doi.org/10.7872/CRYA.V35.ISS2.2014.105

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., Griffiths, R. I., & Schonrogge, K. (2015). PIPITS: An automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution*, *6*(8), 973–980. https://doi.org/10.1111/2041-210X.12399

Hanafy, R. A., Johnson, B., Youssef, N. H., & Elshahed, M. S. (2020). Assessing anaerobic gut fungal diversity in herbivores using D1/D2 large ribosomal subunit sequencing and multi-year isolation. *Environmental Microbiology*, *22*(9), 3883–3908. https://doi.org/10.1111/1462-2920.15164;PAGE:STRING:ARTICLE/CHAPTER

Hao, X., Jiang, R., & Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, *27*(5), 611–618. https://doi.org/10.1093/BIOINFORMATICS/BTQ725

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., Janzen, D. H., Hallwachs, W., Naik, S., Sones, J. E., & Zakharov, E. V. (2018). A Sequel to Sanger: Amplicon sequencing that scales. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4611-3

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., Overmann, J., Mazzoni, C. J., & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Molecular Ecology Resources*, *18*(6), 1500–1514. https://doi.org/10.1111/1755-0998.12937

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., & Raes, J. (2014). LotuS: An efficient and user-friendly OTU processing pipeline. *Microbiome*, *2*(1), 1–7. https://doi.org/10.1186/2049-2618-2-30/TABLES/3

Hu, Y., Irinyi, L., Hoang, M. T. V., Eenjes, T., Graetz, A., Stone, E. A., Meyer, W., Schwessinger, B., & Rathjen, J. P. (2022). Inferring Species Compositions of Complex Fungal Communities from Long- and Short-Read Sequence Data. *MBio*, *13*(2). https://doi.org/10.1128/MBIO.02444-21/SUPPL_FILE/MBIO.02444-21-S0002.PPTX

Hupfauf, S., Etemadi, M., Juárez, M. F. D., Gómez-Brandón, M., Insam, H., & Podmirseg, S. M. (2020). CoMA – an intuitive and user-friendly pipeline for amplicon-sequencing data analysis. *PLoS ONE*, *15*(12 December). https://doi.org/10.1371/journal.pone.0243241

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1103-0

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., & Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution 2022 6:10*, *6*(10), 1458–1470. https://doi.org/10.1038/s41559-022-01838-4

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., & Burki, F. (2020). Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, *20*(2), 429–443. https://doi.org/10.1111/1755-0998.13117

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*, *18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/MOLBEV/MST010

Kebschull, J. M., & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, *43*(21). https://doi.org/10.1093/nar/gkv717

Köster, J., Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., & Nahnsen, S. (2021). Sustainable data analysis with Snakemake. *F1000Research 2021 10:33*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, C., Sieracki, M. E., Jaillon, O., Wincker, P., Vandepoele, K., Logares, R., & Massana, R. (2021). Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal*, *15*(6), 1767–1781. https://doi.org/10.1038/S41396-020-00885-8

Latz, M. A. C., Grujcic, V., Brugel, S., Lycken, J., John, U., Karlson, B., Andersson, A., & Andersson, A. F. (2022). Short- and long-read metabarcoding of the eukaryotic rRNA operon: Evaluation of primers and comparison to shotgun metagenomics sequencing. *Molecular Ecology Resources*, *22*(6), 2304–2318. https://doi.org/10.1111/1755-0998.13623;PAGEGROUP:STRING:PUBLICATION

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., Cruickshank, R., Dopheide, A., Handley, K. M., Hermans, S., Kamke, J., Lee, C. K., Macdiarmid, R., Morales, S. E., Orlovich, D. A., Smissen, R., Wood, J., & Holdaway, R. (2018). Methods for the extraction, storage, amplification and sequencing of dna from environmental samples. *New Zealand Journal of Ecology*, *42*(1). https://doi.org/10.20417/nzjecol.42.9

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/BIOINFORMATICS/BTL158

Lindgreen, S. (2012). AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Research Notes*, *5*(1), 1–7. https://doi.org/10.1186/1756-0500-5-337/COMMENTS

Liu, S., Yang, C., Zhou, C., & Zhou, X. (2017). Filling reference gaps via assembling DNA barcodes using high-throughput sequencing-Moving toward barcoding the world. In *GigaScience* (Vol. 6, Issue 12, pp. 1–8). Oxford University Press. https://doi.org/10.1093/gigascience/gix104

Loos, D., Zhang, L., Beemelmanns, C., Kurzai, O., & Panagiotou, G. (2021). DAnIEL: A User-Friendly Web Server for Fungal ITS Amplicon Sequencing Data. *Frontiers in Microbiology*, *12*, 720513. https://doi.org/10.3389/FMICB.2021.720513/BIBTEX

Lu, J., Zhang, X., Zhang, X., Wang, L., Zhao, R., Liu, X. Y., Liu, X., Zhuang, W., Chen, L., Cai, L., & Wang, J. (2023). Nanopore sequencing of full rRNA operon improves resolution in mycobiome analysis and reveals high diversity in both human gut and environments. *Molecular Ecology*, *32*(23), 6330–6344. https://doi.org/10.1111/MEC.16534

Lu, N., Qiao, Y., An, P., Luo, J., Bi, C., Li, M., Lu, Z., & Tu, J. (2023). Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics*, *24*(5), 1–11. https://doi.org/10.1093/BIB/BBAD275

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560–564. https://doi.org/10.1073/PNAS.74.2.560

Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwälder, B., Schlötterer, C., & Boenigk, J. (2010). Diversity in a hidden world: Potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, *19*(SUPPL. 1), 32–40. https://doi.org/10.1111/j.1365-294X.2009.04478.x

Mikryukov, V., Anslan, S., & Tedersoo, L. (2025). *NextITS: a pipeline for metabarcoding eukaryotes with full-length ITS sequenced with PacBio*. https://doi.org/10.5281/zenodo.15074881

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/MOLBEV/MSAA015

Mosher, J. J., Bernberg, E. L., Shevchenko, O., Kan, J., & Kaplan, L. A. (2013). Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *Journal of Microbiological Methods*, *95*(2), 175–181. https://doi.org/10.1016/J.MIMET.2013.08.009

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/BIOINFORMATICS/BTT509

Nilsson, R. H., Abarenkov, K., Veldre, V., Nylinder, S., De Wit, P., Brosche, S., Alfredsson, J. F., Ryberg, M., & Kristiansson, E. (2010). An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources*, *10*(6), 1076–1081. https://doi.org/10.1111/j.1755-0998.2010.02850.x

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Borman, T., Carvalho, G., Chirico, M., De Caceres, M., … Weedon, J. (2024). *vegan: Community Ecology Package*. https://CRAN.R-project.org/package=vegan

Öpik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J. M., Reier, Ü., & Zobel, M. (2010). The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, *188*(1), 223–241. https://doi.org/10.1111/J.1469-8137.2010.03334.X;CSUBTYPE:STRING:SPECIAL;PAGE:STRING:ARTICLE/CHAPTER

Overgaard, C. K., Jamy, M., Radutoiu, S., Burki, F., & Dueholm, M. K. D. (2024). Benchmarking long-read sequencing strategies for obtaining ASV-resolved rRNA operons from environmental microeukaryotes. *Molecular Ecology Resources*, *24*(7), e13991. https://doi.org/10.1111/1755-0998.13991;PAGE:STRING:ARTICLE/CHAPTER

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, *2018*(5), e4925. https://doi.org/10.7717/PEERJ.4925/SUPP-1

Porras-Alfaro, A., Liu, K. L., Kuske, C. R., & Xiec, G. (2014). From genus to phylum: Large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. *Applied and Environmental Microbiology*, *80*(3), 829–840. https://doi.org/10.1128/AEM.02894-13/SUPPL_FILE/ZAM999105073SO1.PDF

Porter, T. M., & Hajibabaei, M. (2022). MetaWorks: A flexible, scalable bioinformatic pipeline for high-throughput multi-marker biodiversity assessments. *PLOS ONE*, *17*(9), e0274260. https://doi.org/10.1371/JOURNAL.PONE.0274260

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLOS ONE*, *8*(7), e66213. https://doi.org/10.1371/JOURNAL.PONE.0066213

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. In *Genomics, Proteomics and Bioinformatics* (Vol. 13, Issue 5, pp. 278–289). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2015.08.002

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *2016*(10), e2584. https://doi.org/10.7717/PEERJ.2584/FIG-7

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)* (Vol. 74, Issue 12).

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. https://doi.org/10.1111/1755-0998.12236

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., & Burki, F. (2019). New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution*, *36*(4), 757–765. https://doi.org/10.1093/MOLBEV/MSZ012

Sundh, J., Granqvist, E., Iwaszkiewicz-Eggebrecht, E., Manoharan, L., Dijk, L. J. A. van, Goodsell, R., Godeiro, N. N., Bellini, B. C., Łukasik, P., Miraldo, A., Roslin, T., Tack, A. J. M., Andersson, A. F., & Ronquist, F. (2024). HAPP: High-Accuracy Pipeline for Processing deep metabarcoding data. *BioRxiv*, 2024.12.20.629441. https://doi.org/10.1101/2024.12.20.629441

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. In *Molecular Ecology* (Vol. 21, Issue 8, pp. 1789–1793). https://doi.org/10.1111/j.1365-294X.2012.05542.x

Tedersoo, L., Albertsen, M., Anslan, S., & Callahan, B. (2021). Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Applied and Environmental Microbiology*, *87*(17), 1–19. https://doi.org/10.1128/AEM.00626-21

Tedersoo, L., Anslan, S., Bahram, M., Kõljalg, U., & Abarenkov, K. (2020). Identifying the 'unidentified' fungi: a global-scale long-read third-generation sequencing approach. *Fungal Diversity*, *103*(1), 273–293. https://doi.org/10.1007/S13225-020-00456-4/TABLES/2

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., Anslan, S., & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. In *Molecular Ecology* (Vol. 31, Issue 10, pp. 2769–2795). John Wiley and Sons Inc. https://doi.org/10.1111/mec.16460

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. In *Molecular Ecology Resources* (Vol. 19, Issue 1, pp. 47–76). Blackwell Publishing Ltd. https://doi.org/10.1111/1755-0998.12959

Tedersoo, L., Mikryukov, V., Anslan, S., Bahram, M., Khalid, A. N., Corrales, A., Agan, A., Vasco-Palacios, A. M., Saitta, A., Antonelli, A., Rinaldi, A. C., Verbeken, A., Sulistyo, B. P., Tamgnoue, B., Furneaux, B., Ritter, C. D., Nyamukondiwa, C., Sharp, C., Marín, C., … Abarenkov, K. (2021). The Global Soil Mycobiome consortium dataset for boosting fungal diversity research. *Fungal Diversity*, *111*(1), 573–588. https://doi.org/10.1007/S13225-021-00493-7/FIGURES/6

Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*, *217*(3), 1370–1385. https://doi.org/10.1111/nph.14776

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. In *Biological Conservation* (Vol. 183, pp. 4–18). Elsevier Ltd. https://doi.org/10.1016/j.biocon.2014.11.019

Tikhonenkov, D. V., Mikhailov, K. V., Gawryluk, R. M. R., Belyaev, A. O., Mathur, V., Karpov, S. A., Zagumyonnyi, D. G., Borodina, A. S., Prokina, K. I., Mylnikov, A. P., Aleoshin, V. V., & Keeling, P. J. (2022). Microbial predators form a new supergroup of eukaryotes. *Nature 2022 612:7941*, *612*(7941), 714–719. https://doi.org/10.1038/s41586-022-05511-5

Torruella, G., De Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., Del Campo, J., Eme, L., Pérez-Cordón, G., Whipps, C. M., Nichols, K. M., Paley, R., Roger, A. J., Sitjà-Bobadilla, A., Donachie, S., & Ruiz-Trillo, I. (2015). Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology : CB*, *25*(18), 2404–2410. https://doi.org/10.1016/J.CUB.2015.07.053

Vasar, M., Davison, J., Neuenkamp, L., Sepp, S. K., Young, J. P. W., Moora, M., & Öpik, M. (2021). User-friendly bioinformatics pipeline gDAT (graphical downstream analysis tool) for analysing rDNA sequences. *Molecular Ecology Resources*, *21*(4), 1380–1392. https://doi.org/10.1111/1755-0998.13340

Větrovský, T., Baldrian, P., & Morais, D. (2018). SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. *Bioinformatics*, *34*(13), 2292–2294. https://doi.org/10.1093/BIOINFORMATICS/BTY071

Větrovský, T., Kolařík ová, Z., Lepinay, C., Awokunle Hollá, S., Davison, J., Fleyberková, A., Gromyko, A., Jelínková, B., Kolařík, M., Krüger, M., Lejsková, R., Michalčíková, L., Michalová, T., Moora, M., Moravcová, A., Moulíková, Š., Odriozola, I., Öpik, M., Pappová, M., … Kohout, P. (2023). GlobalAMFungi: a global database of arbuscular mycorrhizal fungal occurrences from high-throughput sequencing metabarcoding studies. *New Phytologist*, *240*(5), 2151–2163. https://doi.org/10.1111/NPH.19283;WGROUP:STRING:PUBLICATION

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07/SUPPL_FILE/SUMMARY_BYHIERARCHY.ZIP

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wijayawardene, N. N., Hyde, K. D., Al-Ani, L. K. T., Tedersoo, L., Haelewaters, D., Rajeshkumar, K. C., Zhao, R. L., Aptroot, A., Leontyev, D. V., Saxena, R. K., Tokarev, Y. S., Dai, D. Q., Letcher, P. M., Stephenson, S. L., Ertz, D., Lumbsch, H. T., Kukwa, M., Issi, I. V., Madrid, H., … Thines, M. (2020). Outline of Fungi and fungus-like taxa. *Mycosphere*, *11*(1), 1060–1456. https://doi.org/10.5943/mycosphere/11/1/8

Yang, C., Zheng, Y., Tan, S., Meng, G., Rao, W., Yang, C., Bourne, D. G., O'Brien, P. A., Xu, J., Liao, S., Chen, A., Chen, X., Jia, X., Zhang, A. bing, & Liu, S. (2020). Efficient COI barcoding using high throughput single-end 400 bp sequencing. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07255-w

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research*, *42*(D1), D643–D648. https://doi.org/10.1093/NAR/GKT1209

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. https://doi.org/10.1111/J.2041-210X.2012.00198.X

SUMMARY IN ENGLISH

Molecular techniques, particularly DNA metabarcoding using high-throughput sequencing (HTS), have emerged as powerful tools, enabling the simultaneous identification of multiple taxa from environmental DNA (eDNA). The ongoing loss of Earth's biodiversity represents one of the most pressing challenges of the century, and traditional methods such as morpho-taxonomy for species communities monitoring are often time-consuming, limited in scope, and prone to inconsistencies due to varying taxonomic expertise. Moreover, ambiguous species and juvenile life stages are frequently unidentifiable to lower taxonomic levels. The advent of molecular methods, especially DNA sequencing technologies, has revolutionized biodiversity assessments. This began with techniques like the Maxam–Gilbert method (in the 1970s) and, more prominently, Sanger sequencing, which became widely used due to its technical simplicity, ability to produce longer sequences, and amenability to automation. Although Sanger sequencing is essential for DNA barcoding of individual specimens, it is unsuitable for characterizing species communities in mixed DNA from environmental samples without tedious cloning steps.

Newer HTS technologies can sequence thousands to millions of DNA fragments simultaneously, making them suitable for metabarcoding workflows. Metabarcoding combines DNA barcoding with HTS to identify multiple species from mixed environmental samples, such as soil, water, or bulk organism collections, using standardized genetic markers. It enables rapid, cost-effective, and non-invasive detection of a broad range of taxa, including elusive or morphologically cryptic species, providing a comprehensive view of community composition. Practical guidelines for metabarcoding have enhanced the scalability and throughput of eDNA sample processing, increasing its appeal among ecologists. However, second-generation HTS platforms such as Illumina are limited by short read lengths, with a maximum of 2 × 300 bp in paired-end mode (and 2 × 500 bp since October 2025), compromising taxonomic resolution and primer design. Short reads may limit species-level resolution, provide limited phylogenetic depth, and make designing suitable target-specific primer binding sites challenging. Standard barcoding regions are generally longer than 500 bp, such as the ~650 bp COI gene for animals, ~1.5 kb 16S rRNA gene for bacteria, or 600–1000+ bp ITS region for fungi.

In the first wave of metabarcoding studies (~2007–2011), Roche's 454 pyrosequencing platform was widely adopted, evolving to support reads of up to 700–1,000 bp, long enough to capture substantial portions of these barcodes. However, it was supplanted by Illumina (in 2011, with the release of MiSeq platform) due to higher throughput, lower costs, and superior accuracy. Illumina's short reads often require mini-barcodes, which can reduce accuracy and necessitate alternative long-read platforms. There has been growing interest in third-generation HTS methods, such as those from Pacific Biosciences (PacBio) and Oxford Nanopore Technology, which offer longer reads that span full-length DNA

markers (e.g., the rRNA operon including SSU, ITS, and LSU), improving phylogenetic depth and accuracy. Regions of the rRNA operon are widely used markers for fungal and eukaryotic barcoding due to their universal presence, multi-copy nature, and balanced variability. The operon comprises coding regions for the SSU and LSU ribosomal RNAs, interspersed with the more variable ITS regions. While ITS is the standard barcode for fungi, SSU and LSU are applied for certain taxonomic groups or when ITS provides insufficient variation. Sequencing the full-length rRNA (SSU-ITS-LSU) operons using long-read technologies enhances taxonomic resolution by reducing biases inherent to single markers.

For PacBio, errors are mitigated by forming circular consensus sequences (CCS), now known as HiFi reads, where an amplified locus is circularized and sequenced multiple times to achieve high accuracy comparable to short amplicons (99.9% per-base accuracy). Despite these advances, long-read data are prone to artifacts, such as chimeras and sequencing errors, which require specialized bioinformatics pipelines for quality control, feature generation (e.g., OTUs or ASVs), and taxonomic annotation. The rapid advancement of HTS platforms has led to a proliferation of software for metabarcoding data processing. A large amount of sequencing data per sample necessitates efficient transformation of raw reads into biodiversity data through sequence analysis pipelines, which apply steps to generate features tables annotated with taxonomic information. Features can be operational taxonomic units (OTUs) from clustering or amplicon sequence variants (ASVs) from denoising. Foundational software, such as mothur, USEARCH, and QIIME, has been augmented with algorithms to minimize artifacts and implement clustering and denoising. Meanwhile, workflow tools like PipeCraft2 wrap these and other bioinformatics software to streamline analysis pipelines.

However, this shift towards long reads comes with several challenges: the sheer number of available analysis pipelines is overwhelming, many of which are not optimized specifically to handle long-read specific errors; there is a lack of curated, full-length reference databases for accurate taxonomic assignment; and long amplicons are more susceptible to artifacts like chimeras, yet the performance of existing detection tools on this data type is poorly understood.

Accordingly, this thesis reviews 32 bioinformatics pipelines, classifying them by workflow structure (pre-complied & software suite), user interface (CLI & GUI), and feature generation methods (clustering for OTUs/swarm-clusters and denoising for ASVs) (I). Here, the EUKARYOME database is introduced as the first curated, multi-marker reference database for all eukaryotes containing full-length rRNA operon sequences, improving taxonomic identification accuracy and chimera validation for long-read data (II). Furthermore, I developed the Snake-EUK pipeline to facilitate the usage of the database by other downstream tools. I designed and performed simulations and empirical data analyses to benchmark *de novo* chimera detection algorithms (*uchime_denovo, chimeras_denovo, removeBimeraDenovo*) on PacBio HiFi long-read datasets, evaluating

performance metrics such as F1 scores and false-positive rates, as well as the impact of parameter tuning and secondary validation strategies (III).

The main results and conclusions of the thesis are as follows: 1) A comprehensive overview of 32 bioinformatics pipelines revealed significant diversity in workflow structures, feature generation approaches (OTUs *vs*. ASVs), and marker specializations, where the selection of an appropriate one depends on the structure of the dataset subject to analysis. Software suites like QIIME 2 and VSEARCH offer greater flexibility compared to precompiled pipelines but also require more advanced bioinformatic skills to apply. Software packages, such as PipeCraft2, that wrap multiple bioinformatic tools but also pre-compiled pipelines, offer a user-friendly solution that bridges the gap between flexibility and accessibility, enabling researchers without extensive bioinformatic expertise to conduct sophisticated analyses. The overview provides a practical guide for selecting appropriate metabarcoding analysis tools depending on the user's data structure and aims of the analyses. 2) The EUKARYOME database, the first curated reference database of long rRNA markers for all eukaryotes, covers over 172,000 species and enables more accurate taxonomic identification and chimera detection compared to existing databases. 3) Evaluation of *de novo* chimera detection algorithms on long-read data reveals high false-positive rates and inconsistent performance, with *uchime_denovo* being the most precise. Accurate diversity estimates require parameter tuning and a secondary validation workflow to effectively remove artifacts while minimizing the loss of true biological data. Parameter tuning and secondary validation workflows effectively reduce artifacts while conserving genuine biological sequences. However, their influence on large-scale community structure was minimal, suggesting that biases inherent to the methods are more significant in taxonomic composition than in overall diversity patterns. These findings and resources collectively advance long-read metabarcoding as a robust tool for biodiversity assessment.

SUMMARY IN ESTONIAN

Pikkade DNA markerjärjestuste masstriipkoodimine: tööriistad ja referentsandmebaasid

Pärilikkusaine (DNA) järjestustehnoloogiate areng on revolutsiooniliselt hõlbustanud organismide tuvastamise võimalusi ning seeläbi ka elurikkuse hindamist. DNA abil organismide määramine põhineb liikidele iseloomulike DNA järjestuste sekveneerimisel ja seejärel andmebaasidesse salvestatud liikide järjestustega võrdlemisel. Traditsioonilised meetodid, näiteks liikide määramine morfoloogia alusel, on koosluste monitooringus sageli väga aeganõudvad, seda eriti mikroskoopiliste organismirühmade puhul. Seetõttu on DNA masstriipkoodistamise (*metabarcoding*) meetod on kujunenud populaarseks organismide ja nende koosluste võrdlemisi kiireks tuvastamiseks.

DNA sekveneerimine sai alguse 1970-ndatel aastatel selliste tehnikatega nagu Maxam–Gilbert'i ja Sangeri meetod. Viimane sai laialdaselt kasutatavaks ja on kasutuses ka tänapäeval oma tehnilise lihtsuse ja automatiseerituse tõttu. Kuigi Sangeri meetodi abil DNA järjestamine on oluline üksikute isendite DNA triipkoodistamiseks ja määramiseks, ei sobi see koosluste kirjeldamiseks keskkonna DNA-st ilma aeganõudva kloonimisetappideta. Uuemad DNA sekveneerimise tehnoloogiad suudavad järjestada miljoneid DNA fragmente samaaegselt, muutes need sobivaks masstriipkoodistamise jaoks. Masstriipkoodistamine ühendab DNA triipkoodistamise ja mass-sekveneerimise, et tuvastada samaaegselt mitmeid liike keskkonnaproovidest, nagu muld, vesi või organismide DNA kogumikud, kasutades standardiseeritud geneetilisi markereid. See võimaldab suhteliselt kiiret ja kulutõhusat liikide koosluste määramist, sealhulgas ka morfoloogiliselt raskesti eristatavate ja krüptiliste liikide (morfoloogiliselt sarnased kui bioloogiliselt erinevad liigid) tuvastamist. Lisaks sekveneerimise tehnoloogiate arengule on viimastel aastatel ilmunud masstriipkoodistamise kasutamise juhendid, mis kirjeldavad protseduure ja parimaid tavasid andmete kogumiseks, analüüsimiseks ja tõlgendamiseks. Need juhendid on suurendanud meetodi ligipääsetavust, muutes masstriipkoodistamise ökoloogide seas laialdasemalt kasutatavaks ja hõlbustades selle rakendamist erinevates uurimisvaldkondades.

Standardseid triipkoodistamise DNA järjestused on üldjuhul pikemad kui 500 aluspaari, näiteks ligikaudu 650 aluspaari tsütokroom c oksüdaasi allüksus I (COI) geeni puhul (loomade määramiseks), umbes 1500 aluspaari 16S ribosomaalse RNA (rRNA) geeni (bakterite määramiseks) ja 600–1000+ aluspaari ribosomaalse DNA sisemise transkribeeritava speisseri (ehk ITS; seente määramiseks) puhul. Teise põlvkonna mass-sekveneerimise platvormid, nagu Illumina, suudavad aga sekveneerida ainult suhteliselt lühikesi fragmente – paarislugemise režiimis maksimaalselt 2 × 300 aluspaari (ja alates oktoobrist 2025 kuni 2 × 500 aluspaari). Lühikesed DNA järjestused ei pruugi pakkuda liigitasandi eristust ning limiteeritud sekveneerimispikkuse tõttu on uute DNA praimerite disaini keerulisem.

Esimeste masstriipkoodistamise meetodit kasutavate uuringute laine ajal (~2007–2011) kasutati laialdaselt Roche 454 pürosekveneerimise platvormi, mis edasiste arengute käigus võimaldades sekveneerida 700–1000 aluspaari pikkuseid DNA fragmente. Need järjestused on piisavalt pikad, et hõlmata märkimisväärset osa täispikast standardsest triipkoodist. Aastal 2011 asendas Illumina MiSeq platvorm suuresti pürosekveneerimise, kuna see võimaldas suuremat sekventside saagist, paremat sekveneerimistäpsust ning madalamaid kulusid. Siiski võimaldas Illumina platvorm sekveneerida maksimaalselt ainult 2 × 300 aluspaari (ehk ~550 aluspaari kui mõlemad paarislugemise režiimi sekventsid usaldusväärselt kokku siduda). Seetõttu kasutatakse masstriipkoodistamise töödes sageli niinimetatud mini-triipkoode, mis on täispika triipkoodi lühendatud versioonid.

Viimasel aastakümnel on kolmanda põlvkonna mass-sekveneerimise meetodite täpsus oluliselt paranenud ning platvormid nagu Pacific Biosciences (PacBio) ja Oxford Nanopore Technology võimaldavad muuhulgas sekveneerida ka täispikkasid DNA markergeene (triipkoode). Need platvormid võimaldavad järjestada näiteks täispikki ribosomaalse DNA ITS regioone koos külgnevate geenidega, mida kasutatakse laialdaselt seente ja teiste eukarüootide määramisel. PacBio sekveneerimise käigus loetakse ühte molekuli mitmeid kordi, et moodustada väga täpne konsensusjärjestus, mis on võrreldav lühikeste (Illumina) järjestustega (aluspaari täpsus 99,9%). Seetõttu on nüüd võimalik taas pöörduda täispikkade triipkoodide sekveneerimise poole, et ära kasutada pikemate järjestuste paremat täpsust liikide määramisel ning lähedalt suguluses olevate liikide eristamisel.

Siiski toob üleminek pikkadele järjestustele kaasa mitmeid väljakutseid: paljud olemasolevad sekveneerimise andmete töötlemise tarkvarad ei ole optimeeritud pikkadele järjestustele ning viimastele puuduvad ka kureeritud suuremahulised referentsandmebaasid. Pikemad amplikonid (polümeraasi ahelreaktsiooni käigus sünteesitud nukleotiidijärjestused) võivad olla ka vastuvõtlikumad järjestuste tekkimisele, kus kaks või enam erinevat DNA fragmenti ühinevad üheks järjestuseks (kimäärsed järjestused). Kimäärid on probleemsed, sest need kujutavad endast kunstlikke hübriide, mis ei esine looduses, ning võivad seega vähendavad andmete usaldusväärsust. Samas ei ole olemasolevate kimääride tuvastamise tööriistade efektiivsust pikemate järjestuste puhul põhjalikult uuritud. Suur hulk sekveneerimise andmeid nõuab efektiivset toorandmete teisendamist elurikkuse andmeteks läbi bioinformaatiliste töövoogude. Mass-sekveneerimise platvormide kiire areng on toonud kaasa ka masstriipkoodistamise andmete töötlemiseks mõeldud tarkvara kiire arengu. Esimesed mass-sekveneerimise andmete analüüsimise tarkvara-paketid nagu mothur, USEARCH ja QIIME on järjepidevalt täiendatud algoritmidega, millede rakendamine aitavad välja filtreerida sekveneerimise käigus tekkinud artefakte, samas kui tarkvarad nagu PipeCraft2 ühendavad neid ja teisi bioinformaatika tööriistu, et lihtsustada analüüside töövooge.

Mitmekümnetes masstriipkoodistamise jaoks arendatud bioinformaatika tarkvarades orienteerumiseks, teeb käesolev doktoritöö enim kasutatavatest tarkvarast

põhjaliku ülevaate (I). Järgnevalt tutvustab doktoritöö esimest kureeritud, mitme markeri referentsandmebaasi, EUKARYOME, mis sisaldab täispikkasid rRNA operoni järjestusi, parandades taksonoomilise määramise täpsust ja kimääride valideerimist pikkade järjestuste andmestikes (II). Lisaks on antud doktoritöö käigus välja töötatud SnakeEUK töövoog, mis hõlbustab EUKARYOME andmebaasi kasutamist järjestuste määramiseks mitmete tööriistade kaudu. Viimaks käsitleb antud doktoritöö põhjalikku analüüsi olemasolevate kimääride tuvastamise algoritmide efektiivsusest kimääride filtreerimises pikkade järjestuste andmestikes. Selle tarbeks on läbi viidud simulatsioone ning empiirilisi andmeanalüüse, et hinnata kimääride tuvastamise algoritmide, nagu *uchime_denovo*, *chimeras_denovo* ja *removeBimeraDenovo*, spetsiifilisust ja täpsust PacBio platvormil sekveneeritud täispikkadel ITS järjestustel (III).

Töö peamised tulemused ja järeldused on järgmised: 1) bioinformaatika tarkvarade põhjalik ülevaade näitas suurt mitmekesisust andmete töötlemise strateegia osas, kus sobiliku tarkvara valik sõltub suuresti analüüsitavast andmestruktuurist. Tarkvarapaketid nagu QIIME 2 ja VSEARCH pakuvad suuremat analüüsimeetodite paindlikust võrreldes eelkomplekteeritud bioinformaatiliste töövoogudega, kuid nende kasutamine nõuab kasutajalt ka rohkem bioinformaalilisemaid oskusi. Tarkvarapaketid, mis ühendavad mitmeid bioinformaatika tööriistu ja eelkomplekteeritud töövooge (nagu PipeCraft2), pakuvad kasutajasõbralikku lahendust, ühendades paindlikkuse ja kättesaadavuse ning võimaldades ka ulatuslike bioinformaatiliste teadmisteta kasutajal läbi viia keerukaid analüüse. Sõltuvalt analüüsitava andmestiku struktuurist ja vajadustest on ülevaates välja toodud sobilikud tööriistad, pakkudes seeläbi juhendit sobivate analüüsivahendite valimiseks. 2) EUKARYOME andmebaas, kui esimene kureeritud pikkade rRNA markerite referentsandmebaas kõikidele eukarüootidele, hõlmab üle 172 000 liigi ning võimaldab täpsemat taksonoomilisi määramist ja kimääride tuvastamist võrreldes olemasolevate andmebaasidega. 3) Kimääride tuvastamine pikkadest järjestustest näitas, et paljud bioloogilised järjestused klassifitseeriti kimääridena ehk valepositiivsetena. Erinevate kimääride tuvastamise algoritmide efektiivsuses esines suur erinevus, kuid nende peenhäälestamine ja sekundaarne kimääride valideerimine vähendas seda. Sekundaarne kimääride valideerimine minimeeris bioloogiliste andmete kadu, kuid ei muutnud uuringu üldisi bioloogilisi mustreid võrreldes andmetega, kus sekundaarset valideerimist ei rakendatud (kimääride filtreerimine tavapäraste seadistustega). Antud doktoritöö tulemused ja ressursid edendavad pikkade järjestuste masstriipkoodistamise töövoogu kui elurikkuse hindamise usaldusväärset tööriista.

ACKNOWLEDGMENTS

This journey has been long, demanding, and deeply rewarding. It would not have been possible without the people whose presence made it meaningful. I am grateful for the opportunity to acknowledge them here.

I owe my deepest gratitude to my supervisors, Sten Anslan and Leho Tedersoo, for their guidance, patience, and trust. Their mentorship shaped my research and helped me grow as a scientist and as a person. My sincere appreciation goes to Vladimir Mikryukov, who first opened the door to the world of pipeline managers and never hesitated to share his time and expertise. I would like to thank my colleagues and friends: Saleh, Farzad, Mahdieh, Daniyal, Israel, Enrico, Eleonora, Meirong, John, Mani, Farah, Muhammad, Yangchunzi, Ervinna, and many others for their companionship, support, and the moments of laughter that made our long research days lighter. To all others who shared this journey with me, your contributions are deeply valued. I thank Toomas Tammaru for the feedback and suggestions to improve the text of this thesis.

A special thank you to my closest friends, Mehran Khodaei and Reza Pashaei, whose steady encouragement and wisdom guided me through many scientific and personal challenges over the past four years.

Finally, I would like to express my heartfelt gratitude to my family for their unwavering love and support from afar, especially to my father. This work is dedicated to you, Dad. You never had the chance to pursue the scientific path you deserved, but your curiosity and spirit live through every page of this thesis. This achievement belongs to both of us.

The research presented in this thesis was supported by the Estonian Science Foundation (grant MOBERC116) and by the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP198).

PUBLICATIONS

CURRICULUM VITAE

Name: Ali Hakimzadeh
Date of birth: 18th of August 1994
Citizenship: Iranian
E-mail: alihakimzadeh73@gmail.com

Education:

2012–2016	Tabriz University, Iran, B.Sc. student, Zoology
2017–2019	University of Siena, Italy, M.Sc. Medical Biotechnologies
2021–	University of Tartu, Estonia, Ph.D. student

Publications:

Ebrahimi, N., Hakimzadeh, A., Bozorgmand, F., Speed, S., Manavi, M. S., Khorram, R., Fahani, K., Rezaei-Tazangi, F., Mansouri, A., Hamblin, M. R., & Aref, A. R. (2023). Role of non-coding RNAs as new therapeutic targets in regulating the EMT and apoptosis in metastatic gastric and colorectal cancers. Cell Cycle, 22(20), 2302–2323. https://doi.org/10.1080/15384101.2023.2286804

Hakimzadeh, A., Asbun, A. A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Durling, M. B., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., Palmer, J. M., . . . Anslan, S. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, *24*(5). https://doi.org/10.1111/1755-0998.13847

Tedersoo, L., Moghaddam, M. S. H., Mikryukov, V., Hakimzadeh, A., Bahram, M., Nilsson, R. H., Yatsiuk, I., Geisen, S., Schwelm, A., Piwosz, K., Prous, M., Sildever, S., Chmolowska, D., Rueckert, S., Skaloud, P., Laas, P., Tines, M., Jung, J., Choi, J. H., . . . Anslan, S. (2024). EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. Database, 2024. https://doi.org/10.1093/database/baae043

Rojas-Castillo, O. A., Hakimzadeh, A., Tedersoo, L., Jacobsen, D., & Kepfer-Rojas, S. (2025). The impact of oil palm plantations and pastures on benthic prokaryotic and fungal communities in tropical streams. Freshwater Biology, 70(1). https://doi.org/10.1111/fwb.14376

Hakimzadeh, A., Mikryukov, V., Metsoja, M., Tedersoo, L., & Anslan, S. (2025). Are we throwing away good data? Evaluation of chimera detection algorithms on long-read amplicons reveals high false-positive rates across algorithms. PeerJ, 13, e20456. https://doi.org/10.7717/peerj.20456

Hagh-Doust, N., Hakimzadeh, A., Kupagme, J. Y., Dierickx, G., Copoț, O., Mikryukov, V., & Tedersoo, L. (2026). Invisible passengers: the diversity and invasion risks of fungi and bacteria transported via footwear of international airport passengers. Biological Invasions, 28(1). https://doi.org/10.1007/s10530-025-03741-y

ELULOOKIRJELDUS

Nimi: Ali Hakimzadeh
Sünniaeg: 18th of August 1994
Kodakondsus: Iraan
E-post: alihakimzadeh73@gmail.com

Hariduskäik:

2012–2016	Tabriz University, Iran, B.Sc. student, Zoology
2017–2019	University of Siena, Italy, M.Sc. Medical Biotechnologies
2021–	University of Tartu, Estonia, Ph.D. student

Publikatsioonid:

Ebrahimi, N., Hakimzadeh, A., Bozorgmand, F., Speed, S., Manavi, M. S., Khorram, R., Fahani, K., Rezaei-Tazangi, F., Mansouri, A., Hamblin, M. R., & Aref, A. R. (2023). Role of non-coding RNAs as new therapeutic targets in regulating the EMT and apoptosis in metastatic gastric and colorectal cancers. Cell Cycle, 22(20), 2302–2323. https://doi.org/10.1080/15384101.2023.2286804

Hakimzadeh, A., Asbun, A. A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Durling, M. B., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., Palmer, J. M., . . . Anslan, S. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, *24*(5). https://doi.org/10.1111/1755-0998.13847

Tedersoo, L., Moghaddam, M. S. H., Mikryukov, V., Hakimzadeh, A., Bahram, M., Nilsson, R. H., Yatsiuk, I., Geisen, S., Schwelm, A., Piwosz, K., Prous, M., Sildever, S., Chmolowska, D., Rueckert, S., Skaloud, P., Laas, P., Tines, M., Jung, J., Choi, J. H., . . . Anslan, S. (2024). EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. Database, 2024. https://doi.org/10.1093/database/baae043

Rojas-Castillo, O. A., Hakimzadeh, A., Tedersoo, L., Jacobsen, D., & Kepfer-Rojas, S. (2025). The impact of oil palm plantations and pastures on benthic prokaryotic and fungal communities in tropical streams. Freshwater Biology, 70(1). https://doi.org/10.1111/fwb.14376

Hakimzadeh, A., Mikryukov, V., Metsoja, M., Tedersoo, L., & Anslan, S. (2025). Are we throwing away good data? Evaluation of chimera detection algorithms on long-read amplicons reveals high false-positive rates across algorithms. PeerJ, 13, e20456. https://doi.org/10.7717/peerj.20456

Hagh-Doust, N., Hakimzadeh, A., Kupagme, J. Y., Dierickx, G., Copoț, O., Mikryukov, V., & Tedersoo, L. (2026). Invisible passengers: the diversity and invasion risks of fungi and bacteria transported via footwear of international airport passengers. Biological Invasions, 28(1). https://doi.org/10.1007/s10530-025-03741-y

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

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

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

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.
62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.

63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.
83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.

84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.

106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **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 intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.

206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.

225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.

246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.

265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p
282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.

284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.
285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.
301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.

303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.
321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.

324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.
325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik**. Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak**. Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak**. Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar**. Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo**. Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova**. Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare**. The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.
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.
369. **Martti Vasar**. Developing a bioinformatics pipeline gDAT to analyse arbuscular mycorrhizal fungal communities using sequence data from different marker regions. Tartu, 2020, 193 p.
370. **Ott Kangur**. Nocturnal water relations and predawn water potential disequilibrium in temperate deciduous tree species. Tartu, 2020, 126 p.
371. **Helen Post**. Overview of the phylogeny and phylogeography of the Y-chromosomal haplogroup N in northern Eurasia and case studies of two linguistically exceptional populations of Europe – Hungarians and Kalmyks. Tartu, 2020, 143 p.
372. **Kristi Krebs**. Exploring the genetics of adverse events in pharmacotherapy using Biobanks and Electronic Health Records. Tartu, 2020, 151 p.
373. **Kärt Ukkivi**. Mutagenic effect of transcription and transcription-coupled repair factors in *Pseudomonas putida.* Tartu, 2020, 154 p.
374. **Elin Soomets**. Focal species in wetland restoration. Tartu, 2020, 137 p.
375. **Kadi Tilk**. Signals and responses of ColRS two-component system in *Pseudomonas putida.* Tartu, 2020, 133 p.
376. **Indrek Teino**. Studies on aryl hydrocarbon receptor in the mouse granulosa cell model. Tartu, 2020, 139 p.
377. **Maarja Vaikre**. The impact of forest drainage on macroinvertebrates and amphibians in small waterbodies and opportunities for cost-effective mitigation. Tartu, 2020, 132 p.
378. **Siim-Kaarel Sepp**. Soil eukaryotic community responses to land use and host identity. Tartu, 2020, 222 p.
379. **Eveli Otsing**. Tree species effects on fungal richness and community structure. Tartu, 2020, 152 p.
380. **Mari Pent**. Bacterial communities associated with fungal fruitbodies. Tartu, 2020, 144 p.

381. **Einar Kärgenberg**. Movement patterns of lithophilous migratory fish in free-flowing and fragmented rivers. Tartu, 2020, 167 p.
382. **Antti Matvere**. The studies on aryl hydrocarbon receptor in murine granulosa cells and human embryonic stem cells. Tartu, 2021, 163 p.
383. **Jhonny Capichoni Massante**. Phylogenetic structure of plant communities along environmental gradients: a macroecological and evolutionary approach. Tartu, 2021, 144 p.
384. **Ajai Kumar Pathak**. Delineating genetic ancestries of people of the Indus Valley, Parsis, Indian Jews and Tharu tribe. Tartu, 2021, 197 p.
385. **Tanel Vahter**. Arbuscular mycorrhizal fungal biodiversity for sustainable agroecosystems. Tartu, 2021, 191 p.
386. **Burak Yelmen**. Characterization of ancient Eurasian influences within modern human genomes. Tartu, 2021, 134 p.
387. **Linda Ongaro**. A genomic portrait of American populations. Tartu, 2021, 182 p.
388. **Kairi Raime**. The identification of plant DNA in metagenomic samples. Tartu, 2021, 108 p.
389. **Heli Einberg**. Non-linear and non-stationary relationships in the pelagic ecosystem of the Gulf of Riga (Baltic Sea). Tartu, 2021, 119 p.
390. **Mickaël Mathieu Pihain**. The evolutionary effect of phylogenetic neighbourhoods of trees on their resistance to herbivores and climatic stress. Tartu, 2022, 145 p.
391. **Annika Joy Meitern**. Impact of potassium ion content of xylem sap and of light conditions on the hydraulic properties of trees. Tartu, 2022, 132 p.
392. **Elise Joonas**. Evaluation of metal contaminant hazard on microalgae with environmentally relevant testing strategies. Tartu, 2022, 118 p.
393. **Kreete Lüll**. Investigating the relationships between human microbiome, host factors and female health. Tartu, 2022, 141 p.
394. **Triin Kaasiku**. A wader perspective to Boreal Baltic coastal grasslands: from habitat availability to breeding site selection and nest survival. Tartu, 2022, 141 p.
395. **Meeli Alber**. Impact of elevated atmospheric humidity on the structure of the water transport pathway in deciduous trees. Tartu, 2022, 170 p.
396. **Ludovica Molinaro**. Ancestry deconvolution of Estonian, European and Worldwide genomic layers: a human population genomics excavation. Tartu, 2022, 138 p.
397. **Tina Saupe**. The genetic history of the Mediterranean before the common era: a focus on the Italian Peninsula. Tartu, 2022, 165 p.
398. **Mari-Ann Lind**. Internal constraints on energy processing and their consequences: an integrative study of behaviour, ornaments and digestive health in greenfinches. Tartu, 2022, 137 p.
399. **Markus Valge**. Testing the predictions of life history theory on anthropometric data. Tartu, 2022, 171 p.
400. **Ants Tull**. Domesticated and wild mammals as reservoirs for zoonotic helminth parasites in Estonia. Tartu, 2022, 152 p.

401. **Saleh Rahimlouye Barabi**. Investigation of diazotrophic bacteria association with plants. Tartu, 2022, 137 p.
402. **Farzad Aslani**. Towards revealing the biogeography of belowground diversity. Tartu, 2022, 124 p.
403. **Nele Taba**. Diet, blood metabolites, and health. Tartu, 2022, 163 p.
404. **Katri Pärna**. Improving the personalized prediction of complex traits and diseases: application to type 2 diabetes. Tartu, 2022, 190 p.
405. **Silva Lilleorg**. Bacterial ribosome heterogeneity on the example of bL31 paralogs in *Escherichia coli.* Tartu, 2022, 189 p.
406. **Oliver Aasmets**. The importance of microbiome in human health. Tartu, 2022, 123 p.
407. **Henel Jürgens**. Exploring post-translational modifications of histones in RNA polymerase II-dependent transcription. Tartu, 2022, 147 p.
408. **Mari Tagel**. Finding novel factors affecting the mutation frequency: a case study of tRNA modification enzymes TruA and RluA. Tartu, 2022, 176 p.
409. **Marili Sell**. The impact of environmental change on ecophysiology of hemiboreal tree species – acclimation mechanisms in belowground. Tartu, 2022, 163 p.
410. **Kaarin Hein**. The hissing behaviour of Great Tit (*Parus major*) females reflects behavioural phenotype and breeding success in a wild population. Tartu, 2022, 96 p.
411. **Maret Gerz**. The distribution and role of mycorrhizal symbiosis in plant communities. Tartu, 2022, 206 p.
412. **Kristiina Nõomaa**. Role of invasive species in brackish benthic community structure and biomass changes. Tartu, 2023, 151 p.
413. **Anton Savchenko**. Taxonomic studies in Dacrymycetes: *Cerinomyces* and allied taxa. Tartu, 2023, 181 p.
414. **Ahto Agan**. Interactions between invasive pathogens and resident mycobiome in the foliage of trees. Tartu, 2023, 155 p.
415. **Diego Pires Ferraz Trindade**. Dark diversity dynamics linked to global change: taxonomic and functional perspective. Tartu, 2023, 134 p.
416. **Madli Jõks**. Biodiversity drivers in oceanic archipelagos and habitat fragments, explored by agent-based simulation models. Tartu, 2023, 116 p.
417. **Ciara Baines**. Adaptation to oncogenic pollution and natural cancer defences in the aquatic environment. Tartu, 2023, 164 p.
418. **Rain Inno**. Placental transcriptome and miRNome in normal and complicated pregnancies. Tartu, 2023, 145 p.
419. **Daniyal Gohar**. Diversity, genomics, and potential functions of fungus-inhabiting bacteria. Tartu, 2023, 138 p.
420. **Sirli Rosendahl**. Fitness effects of chromosomal toxin-antitoxin systems in *Pseudomonas putida.* Tartu, 2023, 154 p.
421. **Mathilde Frédérique E. André**. New Guinea, a hotspot for Human evolution: settlement history and adaptation in northern Sahul. Tartu, 2023, 202 p.

422. **Vlad-Julian Piljukov**. Biochemical characterization of Irc3 helicase. Tartu, 2023, 137 p.
423. **Gerli Albert**. Carbon use strategies of macrophyte communities in the northeastern Baltic Sea: implications for a high CO_2 environment. Tartu, 2023, 128 p.
424. **Mariann Koel**. The molecular interactions between trophoblast and endometrial cells in embryo implantation. Tartu, 2023, 171 p.
425. **Robin Gielen**. Diversity and ecological role of pathogenic fungi in insect populations. Tartu, 2023, 139 p.
426. **Kaspar Reier**. Quantity, stability and disparity of ribosomal components in *Escherichia coli* stationary phase. Tartu, 2023, 151 p.
427. **Linda Rusalepp**. The impact of environmental drivers and competition on phenolic metabolite profiles in hybrid aspen and silver birch. Tartu, 2023, 153 p.
428. **Eliisa Pass**. The effect of managed forest-wetland landscapes on forest grouse and nest predation. Tartu, 2023, 115 p.
429. **Sanni Färkkilä**. Methods for studying plant-fungal interactions – reflecting on the old, the new and the upcoming. Tartu, 2024, 147 p.
430. **Maarja Jõeloo**. Advances in microarray-based copy number variation discovery and phenotypic associations. Tartu, 2024, 209 p.
431. **Natàlia Pujol Gualdo**. Decoding genetic associations of female reproductive health traits. Tartu, 2024, 205 p.
432. **Sirelin Sillamaa**. The role of helicases Hmi1 and Irc3 in yeast mitochondrial DNA maintenance. Tartu, 2024, 189 p.
433. **Iris Reinula**. Genetic variation of grassland plants in changing landscapes. Tartu, 2024, 201 p.
434. **Vi Ngan Tran**. The cellular dynamics and epithelial morphogenesis in *Drosophila* wing development. Tartu, 2024, 158 p.
435. **Slendy Julieth Rodríguez Alarcón**. Intraspecific trait diversity in plants: characterizing effects of trait variation on community assembly and ecosystem functioning. Tartu, 2024, 129 p.
436. **Arun Kumar Devarajan**. Microbes and climate change: insights from plant-microbe interactions in rice phyllosphere and soil microbiomes in subarctic grasslands. Tartu, 2024, 224 p.
437. **Leonard Owuraku Opare**. Rearing density effects on a commercially important insect species. Tartu, 2024, 145 p.
438. **Siqiao Liu**. The effect of anthropogenic disturbance on soil fungal communities. Tartu, 2024, 172 p.
439. **Kertu Liis Krigul**. The gut microbiome at the interface of human health and disease. Tartu, 2024, 158 p.
440. **Danat Yermakovich**. The evolutionary history of complex traits: implications of archaic admixture. Tartu, 2024, 153 p.
441. **Yiming Meng**. Plant mycorrhizal type and status in the global flora. Tartu, 2024, 200 p.

442. **Iryna Yatsiuk**. Evolution, species delimitation and diversity in myxomycetes: *Arcyria* and allied genera. Tartu, 2024, 193 p.
443. **Daniela León Velandia**. Mycorrhizal trait distribution and composition in plant communities under natural gradients. Tartu, 2024, 121 p.
444. **Bruno Paganeli**. Dark diversity methods for prioritization of areas and species in nature conservation. Tartu, 2024, 155 p.
445. **Mario Reiman**. Placental transcriptome in normal and complicated pregnancies. Tartu, 2025, 167 p.
446. **Maarja Kõrkjas**. Dynamics of tree-related microhabitats in live forest trees and its links with biodiversity. Tartu, 2025, 134 p.
447. **Eleonora Beccari**. Mapping and exploring trait spaces across the tree of life. Tartu, 2025, 190 p.
448. **Jack R. Hall**. Dissolved organic carbon dynamics of Baltic Sea macroalgae: production, bioavailability and ecosystem effects. Tartu, 2025, 135 p.
449. **Artjom Stepanjuk**. Function of adhesion molecules and signalling pathways in human endometrial and embryonic models. Tartu, 2025, 247 p.
450. **Marianne Kivastik**. Heterostylous plants in an era of global change: the role of local, landscape and climatic actors. Tartu, 2025, 167 p.
451. **Yehor Yatsiuk**. Large tree-cavities as key structures for forest biodiversity. Tartu, 2025, 215 p.
452. **Ovidiu Copoț**. Relevance of eDNA, citizen science, and species distribution modelling for fungal conservation. Tartu, 2025, 198 p.
453. **Tarmo Puurand**. Human genome studies with k-mer frequencies. Tartu, 2025, 184 p.
454. **Stênio Ítalo Araújo Foerster**. Phylogenetic comparative studies of body size in insects and arachnids: from predictions to applications. Tartu, 2025, 171 p.
455. **Hanna Maria Kariis**. Improving pharmacotherapy outcomes in psychiatric and cardiovascular conditions. Tartu, 2025, 193 p.
456. **Elisabeth Prangel**. The impact of land-use change and ecological restoration on biodiversity and ecosystem service supply in semi-natural grasslands. Tartu, 2025, 233 p.
457. **Nidal Fetnassi**. Determinants of moth assemblages across human-modified landscapes of Estonia and Morocco. Tartu, 2025, 162 p.
458. **Vineesh Nedumpally**. Assembling the phylogenetic tree of northern European macroheteroceran moths. Tartu, 2025, 171 p.