DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
388

KAIRI RAIME

The identification of plant DNA in metagenomic samples

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

388

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
388

KAIRI RAIME

The identification of plant DNA in metagenomic samples

UNIVERSITY OF TARTU
Press

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

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Gene Technology on July 5, 2021 by the Council of the Institute of Molecular Cell Biology, University of Tartu.

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

Reviewer: Prof. Ants Kurg, PhD
Chair of Biotechnology, Institute of Molecular and Cell Biology,
University of Tartu, Tartu, Estonia

Opponent: Filipe Pereira, PhD

1. Department of Life Sciences, Centre for Functional Ecology, University of Coimbra, Portugal
2. IDENTIFICA, Science and Technology Park of the University of Porto, Portugal

Commencement: Room No. 105, 23B Riia St., Tartu, on August 25, 2021, at 11:15 am.

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

European Union
European Regional
Development Fund

Investing
in your future

ISSN 1024-6479
ISBN 978-9949-03-681-3 (print)
ISBN 978-9949-03-682-0 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

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

I Kõressaar, T., Lepamets, M., Kaplinski, L., **Raime, K.**, Andreson, R., Remm, M., 2018. Primer3_masker: integrating masking of template sequence with primer design software. *Bioinformatics* 34, 1937–1938. DOI: https://doi.org/10.1093/bioinformatics/bty036.

II **Raime, K.**, Remm, M., 2018. Method for the Identification of Taxon-Specific *k*-mers from Chloroplast Genome: A Case Study on Tomato Plant (Solanum lycopersicum). *Front. Plant Sci.* 9, 6. DOI: https://doi.org/10.3389/fpls.2018.00006.

III **Raime, K.**, Krjutškov, K., Remm, M., 2020. Method for the Identification of Plant DNA in Food Using Alignment-Free Analysis of Sequencing Reads: A Case Study on Lupin. *Front. Plant Sci.* 11, 646. DOI: https://doi.org/10.3389/fpls.2020.00646.

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

My contributions to the listed publications were as follows:

Ref. I Performed the analysis of the masking extent and style of primer3_masker and RepeatMasker in different plant genomes, and participated in the writing of the manuscript;

Ref. II Constructed the *k*-mer detection pipelines, wrote the Python scripts, analyzed the sequence data and wrote the manuscript; participated in the design of the experiments and interpretation of the results of all analysis;

Ref. III Constructed the *k*-mer detection pipelines, wrote the Python scripts, performed the laboratory experiments (incl. sample material collection, DNA extraction, and PCR), analyzed the sequencing data, and wrote the manuscript. Participated in the design of the experiments and interpretation of the results of all analyses.

LIST OF ABBREVIATIONS

PCR	Polymerase chain reaction
k-mer	Substrings of length *k* in a given string (e.g., DNA sequence)
SNP	Single nucleotide polymorphism
Gbp	Gigabase pairs
NGS	Next-generation sequencing
BLAST	The basic local alignment search tool
CPU	Central processing unit
GB	Gigabyte
WGS	Whole genome sequencing
ITS	Internal transcribed spacer

INTRODUCTION

Metagenomic samples, like food, natural medicine products, or environmental samples, contain DNA from very different sources (plants, animals, bacteria, fungi, etc.). The taxonomical identification of DNA from plants or other organisms gives valuable information about the composition and origin of the sample. DNA-based methods offer a good alternative to morphological and chemical methods for the identification of different taxa from degraded metagenomic samples where other methods are often inapplicable.

Different amplification-based methods have been developed and applied to identify plants from metagenomic samples (including polymerase chain reaction – PCR, and barcoding methods) that rely on the amplification of specific genomic regions using taxa-specific primers. The amplification may not occur efficiently enough to identify taxa, because of the nonspecific binding sites of PCR primers, caused by repeated regions in the targeted genome, especially in large plant genomes, or the nonspecific binding sites in the genomes of nontarget taxa. One of the results of my Ph.D. studies is testing a new tool primer3_masker that uses a *k*-mer-based method to detect and mask PCR failure-prone repeated regions in a genome sequence, before PCR primer design. This program has been applied to different plant genomes to analyze the extent of masking and to compare two different masking tools, primer3_masker and GenomeTester.

Sometimes it is impossible to efficiently amplify the specific target region by PCR, particularly from degraded metagenomic samples. Therefore, there is a need for methods that overcome the limitations of the widely used amplification-based methods (PCR and barcoding methods). This dissertation also introduces the *k*-mer based approach to identify plants from sequencing reads of metagenomics samples and provides *k*-mer-based tools to identify taxa-specific *k*-mers from plant plastid genome for the identification of plant taxa directly from sequencing reads.

Whole genome sequencing allows to get more genomic information about the sample than methods based on the amplification of only a few restricted genomic regions. DNA sequences derived from one sequencing run can give genomic information about all different organisms in the sample. It is more cost-effective to use the DNA sequences from the same sequencing run to answer different questions (e.g., the identification DNA from plants, animals, bacteria, viruses, to detect various allergens, pathogens, endangered species, genetically modified organisms).

Short *k*-mer (in length maximum 32 nucleotides) based approach enables fast comparison and identification of DNA sequences. Short *k*-mers are also detectable in whole genome sequencing reads from processed food samples or degraded samples, where the methods that rely on the amplification of the hundred or thousand nucleotides long barcoding region often fail. The key of the *k*-mer-based approach is to identify the appropriate set of short taxa-specific *k*-mers that are used to detect taxon from whole genome sequencing reads. Alignment-free approach allows to bypass some steps that are prone to mistakes (e.g., alignment, assembling the reads).

The first part of this thesis provides an overview of the importance and challenges related to the identification of plants from degraded metagenomic samples and briefly covers the advantages and drawbacks of different DNA-based methods and bioinformatics strategies applied for the identification plants from metagenomic samples.

In the second part of the thesis, I describe our research that has been carried out to develop different methods for the identification of plants from degraded metagenomic samples.

1. REVIEW OF THE LITERATURE

1.1. Identification of plants – importance and challenges

The identification of plant-based ingredients is important in a variety of different areas, including in food authentication, in analyzing natural herbal products (incl. herbal medicines, spice mixes), or cosmetics, where one plant species may be intentionally or unintentionally replaced with another.

One of the main aims of food quality and safety control is to assure that the composition of the product is in compliance with what is claimed and to identify contaminants or counterfeits in raw material, during food processing, and before market placement. This includes, among others, the declared origin (species, geographical or genetic) of food components. During treatment processes, contamination, either accidental or economically motivated, can lead to incongruences between the declared and real composition of the final food products (Bruno et al., 2019). The identification of components of food, natural herbal products, and other plants containing products is important to assure safety for the consumers (Bruno et al., 2019; Gao et al., 2019; Raclariu et al., 2018; Seethapathy et al., 2019; Speranskaya et al., 2018). However, the food authenticity and the discrepancy between the actual composition of food and composition declared by manufacturers is important for all, the average consumer, for large retail chains and the restaurant food industry, to avoid cases of allergic reactions and poisoning (Singtonat and Osathanunkul, 2015). Therefore, food industries are seeking the opportunity to assure their food products labeling compliance and branding, etc. The limitations of existing methods of analysis require the development and application of new approaches to this task (Danezis et al., 2016).

The risks from unexpected or unreported ingredients used in the herbal or food products may range from simple misleading labeling or reduced therapeutic effectiveness of the herbal products to potentially serious allergic or toxic reactions (Chan, 2003; Heubl, 2010; Raclariu et al., 2018; Spencer and Berman, 2003), unreported ingredients can be dangerous for people with various kinds of dietary restrictions (Speranskaya et al., 2018).

Different safety issues have also arisen in the herbal products industry from the intentional or unintentional use of adulterants and admixture with cheaper undeclared ingredients or substitution (Baker et al., 2012; de Boer et al., 2015; Gao et al., 2019; Ghorbani et al., 2017; Little, 2014a; Little and Jeanson, 2013; Steven G. Newmaster et al., 2013; Ouarghidi et al., 2012; Raclariu et al., 2018; Seethapathy et al., 2019; Stoeckle et al., 2011; Wallace et al., 2012). Also, the demand for endangered plant species as ingredients in traditional natural medicines has caused threats to the survival of several endangered plant species, such as agarwood (*Aquilaria* sp.) (Lee et al., 2016). The success of regulating the trading of endangered species is dependent on the ability to identify components and ingredients derived from endangered species (Arulandhu et al., 2017; Seethapathy et al., 2019).

Historically first, morphological and histological methods have been applied to analyze the composition of food products and some of these methods are still in use. For example, in the analysis of plant samples, studies of the structural features of pollen, epidermal cells, and anatomy of sections of plant organs and a comparison with the data of reference literature are performed. The disadvantages of these methods are the need for highly specialized and highly qualified specialists and the duration of the analysis. Also, these approaches often do not make it possible to identify objects with the required accuracy. For example, pollen analysis in many cases allows us to assign an object only to a certain family or genus, whereas identification of species is often necessary (Prosser and Hebert, 2017).

Food and natural herbal products are highly processed and complex mixtures of numerous ingredients making it impossible to detect the distinctive morphological features of the components. Therefore, using morphological, organoleptic, or microscopic features and standard chemical analytical methods for the identification of plant origin components and authentification of raw material is challenging (Gao et al., 2019; Khan and Smillie, 2012; Raclariu et al., 2018; Zhang et al., 2012). There is a need for rapid and cost-effective molecular tools for the analysis of complex and degraded matrixes.

DNA-based methods play an increasing role in food safety control and food adulteration detection. High stability of DNA allows the analysis of highly processed and degraded metagenomics samples (e.g., food, herbal medicine products, and environmental samples) and trace contaminants. Therefore, recent technologically advanced molecular techniques, based on the amplification and/or sequencing of marker DNA regions, can be a useful diagnostic method to identify food species composition (Bruno et al., 2019; Lo and Shaw, 2018).

1.2. DNA-based methods for the identification of plants from metagenomic samples

1.2.1. Targeted PCR-based methods for the identification of plants

Different PCR-based tests have been developed and applied to identify false descriptions or mislabelling of foods. These methods are used to determine the animal components (Fang and Zhang, 2016; Hossain et al., 2017; Taboada et al., 2017), as well as plant ingredients (Röder et al., 2011). There are also several minor methods associated with PCR for identifying plant species, for example, the analysis of the melting curves of the amplified products of marker sequences (Bosmali et al., 2012; Madesis et al., 2012).

PCR-based methods for food authentication depend on the highly specific amplification of DNA fragments by PCR. One significant disadvantage of these PCR-based methods is the lack of universality of methods for detecting different objects and these methods are not suitable for identifying the species composition of food products that contain additives of unknown origin (Speranskaya et al., 2018). To identify many species, it is necessary to develop and optimize an in-house set of primers and probes.

Developments in DNA sequencing have prompted the development of two approaches (DNA barcoding and DNA metabarcoding) that combine PCR and DNA sequencing and have widely used for authentication of herbal and food products' composition (Bruno et al., 2019; de Boer et al., 2015; Raclariu et al., 2017b, 2017a). DNA barcoding method uses Sanger sequencing of short standard DNA regions, known as DNA barcodes, to identify species (Hebert et al., 2003; Raclariu et al., 2018), and is more suitable for single-target identification. DNA barcoding is a widely applied method for molecular identification to solve very different scientific questions in taxonomy, molecular phylogenetics, population genetics, and biogeography (Hajibabaei et al., 2007; Hebert and Gregory, 2005; Valentini et al., 2009), as well as in industry to check adulterations and substitutions in food products (Jaakola et al., 2010), to monitor the authenticity of food product (Cline, 2012; Di Pinto et al., 2016; Raclariu et al., 2018; Wong and Hanner, 2008) or herbal medicine products to identify of botanical ingredient adulterants (Asahina et al., 2010; Chen et al., 2010; Gao et al., 2019; Srirama et al., 2010; Yao et al., 2009), Applications of DNA barcoding also include forensic analysis (Ferri et al., 2009; Miller Coyle et al., 2005), identification of invasive species (Bleeker et al., 2008; Wiel et al., 2009), tracking illegal wildlife collection and trade of flora and fauna (Chen et al., 2008; Eurlings et al., 2013; Gathier et al., 2013; Janjua et al., 2017) and analysis of species diversity in the gut contents of animals (Soininen et al., 2009).

Conventional DNA barcoding faces some practical limitations that restrict the use of this method. DNA barcoding is well supported and validated in the case of the identification of a single ingredient from unprocessed plant material or single biological raw material, not metagenomics samples with degraded DNA (Bruni et al., 2015; de Boer et al., 2015; Galimberti et al., 2015). However, DNA extracted from food could undergo degradation processes due to the intense physico-chemical conditions of industrial treatments. As a consequence, in several processed foods only very short fragments of DNA are available for the analysis and typically the common DNA barcoding analysis fails (Bauer et al., 2003; Novak et al., 2007).

DNA metabarcoding is an approach that combines DNA barcoding with next-generation sequencing (NGS). The main advantage of DNA metabarcoding is that it enables high-throughput multi-taxa identification and provides great potential for species identification from complex and processed samples composed of multiple ingredients such as food supplements, traditional medicines, and food (Taberlet et al., 2012), where the application of other PCR-based analytical methods is often limited (Arulandhu et al., 2017; Raclariu et al., 2018; Staats et al., 2016).

DNA metabarcoding uses universal PCR primers to mass-amplify the sequences of selected DNA barcode region(s) from different species (Fahner et al., 2016; Staats et al., 2016). The obtained DNA barcode sequences are compared to a DNA sequence reference database for taxonomic identification (Fahner et al., 2016; Taberlet et al., 2012). As the amount of data produced per sample is limited to a particular section of DNA, many samples can be sequenced at the same time, data can be produced cheaper, faster, and at a higher throughput as compared with whole genome sequencing (Bayley, 2019; Egan et al., 2012; Ekblom and Galindo, 2011; Pawar et al., 2017).

DNA metabarcoding still has some limitations, similar to those found in DNA barcoding: DNA fragmentation (Gryson et al., 2002; Hrnčírová et al., 2008), DNA amplification biases (Elbrecht and Leese, 2015; Shokralla et al., 2012), the real 'primer universality' (Deagle et al., 2014), the presence of DNA amplification inhibitors (Schrader et al., 2012), the occurrence of (food) materials in trace and low DNA yield (Elbrecht and Leese, 2015), the accidental laboratory contamination when the target DNA is fragmented and of low concentration. All these restraints hamper species detection in food matrices (Bruno et al., 2019). Plant and animal components in processed samples can be highly processed, the isolation of good-quality DNA can be challenging and the amplification of the selected universal DNA barcode regions may not be successful for each species contained in the complex sample, due to DNA degradation or the lack of PCR primer sequence universality.

Using DNA barcoding or DNA metabarcoding, the identification of plant ingredients is based on the presence of the amplifiable DNA, and false-negative results can be expected if the DNA has been degraded or lost during post-harvest processing or manufacturing (de Boer et al., 2015). Also, the result of barcoding analysis can be affected by the occurrence of a bias during the PCR amplification step using "universal" primers. PCR amplification generates a variable number of template–primer mismatches across species, resulting in a final amplified DNA mixture that does not always reflect the original proportion of each species. This may cause the inaccurate estimation of quantities and limits the quantitative potential of DNA metabarcoding (Bista et al., 2018; Bruno et al., 2019; Elbrecht et al., 2018; Leray et al., 2013; Piñol et al., 2019). DNA metabarcoding data can be used for the qualitative detection of taxa, but not for quantitative assessment of species abundance based on the amount of obtained sequence reads (Raclariu et al., 2018; Staats et al., 2016).

The selection of the most suitable barcode region(s) is a key step of the targeted sequencing-based methods. A specific barcode is a fragment of DNA sequence that has a sufficiently high mutation rate to enable species discrimination for the target group of species (Li et al., 2015). Plants have three different types of genomes: nuclear, mitochondrial, and plastid genomes that can be used to find suitable barcoding region(s). Instead of the widely used term "chloroplast genome", I have used the term "plastid genome" in this work that describes the DNA from all types of plastids that may be in plant cells (proplastids, amyloplasts, chromoplasts, chloroplasts, leucoplasts, etioplasts) and have the same genetic information.

The best gene region for barcoding of land plants has been a matter of debate. The mitochondrial gene regions, including the *COI* region that is widely used for the identification of animal taxa, do not sufficiently distinguish plant species, because of the slow evolutionary rate of the plant mitochondrial genome. Therefore, relatively fast-evolving plastid and nuclear genome regions have been proposed as alternative barcodes for plants (CBOL Plant Working Group, 2009; Hollingsworth et al., 2011; Kress and Erickson, 2007; Newmaster et al., 2006).

The initial goal of DNA barcodes was to find a single universal locus for the identification of all plants. Unfortunately, there was no such universal barcode for all plants, especially in the plastid genome where lineage-specific evolution and non-random spatial patterns of substitution can occur (Ahmed et al., 2012; Jiao et al., 2019). Molecular biologists have used different single loci, or combinations of many different loci for the identification of plants (Techen et al., 2014). Different sets of DNA barcodes for plants have been suggested for different cases (i.e., general taxonomic identification of land plants, identification of medicinal plants, etc.), however, none of these are suitable as universal barcodes (Arulandhu et al., 2017; Taylor and Harris, 2012). The most common barcoding regions for plants are *matK, rbcL*, ITS, ITS2, *psbA-trnH, atpF-atpH, ycf5, psbK-I, psbM, trnD, coxI, nad1, trnL-F, rpoB, rpoC1,* and *rps16* (CBOL Plant Working Group, 2009; Hollingsworth et al., 2011; Kress and Erickson, 2007; Newmaster et al., 2006). These regions have a relatively fast evolutionary rate compared to mitochondrial genes and can distinguish the species based on differences in the sequence, and have conserved regions flanking the ends of the DNA sequence for the binding of universal primers. However, none of the traditional single-locus plant DNA barcodes (e.g. plastid genome barcodes *matK*, *rbcL*, *trn-psbA* or nuclear barcode ITS spacer) are not able to discriminate all different plant species and have several problems: lack sufficient sequence variation between closely related taxa, different primer pairs are required to amplify different taxonomic groups, difficulties of amplification or sequencing, too long barcode region for the identification plants from highly processed samples (Chase et al., 2007; Fazekas et al., 2008; Gao et al., 2019; Li et al., 2015).

To solve the problem of low identification efficiency of a single barcode, a multilocus plant barcode, the combinations of two or three loci from plastid or nuclear genome has been suggested (CBOL Plant Working Group, 2009; Chase et al., 2007; Hollingsworth, 2008; Jiao et al., 2019, 2018; Kress and Erickson, 2007; Yu et al., 2017), however, a universal barcode combination for the identification of all plants has not yet identified. The additional increase in the number of barcode regions increases also the amount of work related to PCR and primer design (Li et al., 2015).

PCR amplification of DNA barcode regions is based on universal primer sets (Chen et al., 2014). Finding suitable locations for metabarcoding primers, DNA sequences that are conserved enough to be universal primers can be challenging and the resulting amplified DNA sequence between these primers may be too conserved to differentiate various species. The primer sequences that are specific only to one taxon or a set of species, can amplify a DNA segment that allows for clear differentiation between those species but can be used only on that restricted group (Wilcox et al., 2015). However, for some genera, there may not be a suitable DNA barcode region for the species-level identification (Zhang et al., 2017).

The primers for barcoding can be designed to target different DNA region sizes. The longer the DNA region, the more genetic information and potentially variable positions it can contain to allow for species discrimination. This is the reason why plant DNA barcode regions are generally several hundred base pairs (bp) long (Chen et al., 2010). However, the length of barcode regions may also offer challenges. Long and universal DNA barcodes, that are widely used, are most appropriately applied when testing fresh or minimally processed material (Parveen et al., 2016). In most food or herbal products, due to various processing methods used in manufacturing, the DNA is often highly degraded into fragments of less than 500 nucleotides in length, in some cases even shorter than 100 bp (Parveen et al., 2016). Therefore, the universal long barcodes ranging from 600 to 800 bp, on average, often result in failed amplification. The smaller the DNA fragments are in a sample, the more likely it is that two primer annealing sites are no longer intact on the same strand, amplification during PCR cannot occur and no results will be detected (Parveen et al., 2016; Särkinen et al., 2012). A false-negative result can occur when the target DNA is present but is too highly fragmented for detection, which may lead to the wrong conclusion that the target DNA is absent.

To solve the problem with fragmented DNA, the use of primer combinations that amplify a very small PCR product (<200 bp), a mini-barcode, have been used to identify species in highly processed samples containing heavily degraded DNA (Arulandhu et al., 2017; Coghlan et al., 2012; Parveen et al., 2016). Mini-barcodes have advantages compared to full-length barcodes in the amplification of DNA from processed plant material (Figure 1, (Parveen et al., 2016)). Different studies have shown higher PCR success rates for mini-barcodes with length 150 bp or shorter (compared to full-length barcoding) when amplifying degraded DNA from food, herbal products, or herbarium specimens (Gao et al., 2019; Little, 2014a; Meusnier et al., 2008; Parveen et al., 2016; Taberlet et al., 2007). Särkinen et al. found that shorter amplicons were linked to higher PCR success rates (Särkinen et al., 2012).

However, mini-barcoding is still based on the amplification of specific genomic regions – mini-barcodes. PCR primers can be designed to amplify shorter mini-barcode regions for degraded DNA samples, but such short DNA regions contain less information and their primers are more restrictive, which makes them unsuitable for universal species barcoding (Cheng et al., 2014; Little, 2014b). Also, designing novel primer pairs based on <200 bp sequences and avoiding dimer formation, hairpin formation, and false priming can be challenging (Gao et al., 2019). Short and specific DNA mini-barcode regions can help identify processed products and should ensure reliable PCR amplification results for degraded DNA; however, this technique is limited by its length constraint.

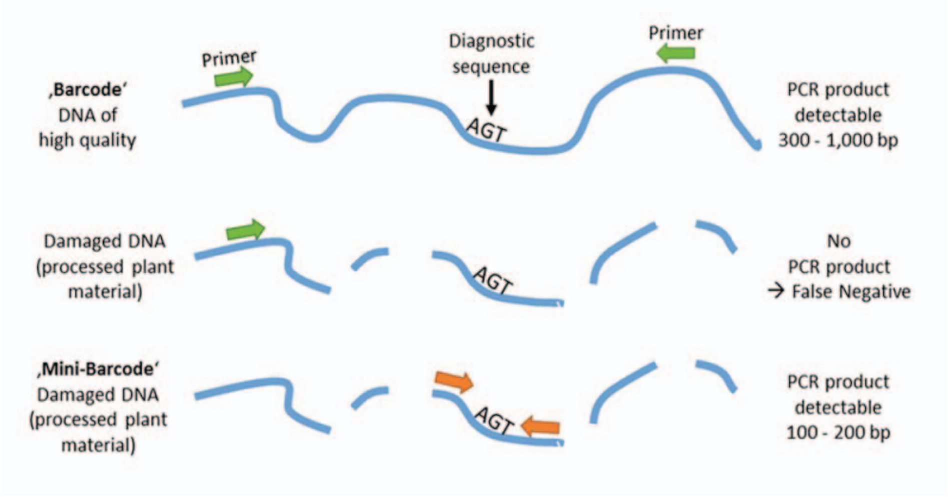
Primer
Diagnostic sequence
Primer
‚Barcode' DNA of high quality
AGT
PCR product detectable 300 - 1,000 bp
Damaged DNA (processed plant material)
AGT
No PCR product → False Negative
‚Mini-Barcode' Damaged DNA (processed plant material)
AGT
PCR product detectable 100 - 200 bp

Figure 1. PCR amplification success of the full-length barcode versus mini-barcode from processed plant material. The product size matters. The amplification of full-length barcode (usually 300–1000 bp) from damaged/fragmented DNA may not be successful (causing false-negative results). Using primers that result in very small PCR product (100–200 bp), a mini-barcode, has shown to have advantages in detecting damaged/fragmented DNA from processed plant materials (Parveen et al., 2016).

DNA quality greatly influences the identification efficiency of different PCR-based methods, including the mini-barcoding method (de Boer et al., 2015; Gao et al., 2019). Both secondary metabolites from plants (fibers, proteins, polysaccharides, polyphenols, alkaloids, etc.) and different ingredients added during processing as preservatives or to improve taste (artificial pigments, starch, salt, tannins, fatty acids, etc.) that are often present in food or natural medicine products reduce the quality of extracted DNA and inhibit PCR amplification (Lo and Shaw, 2018; Madesis et al., 2014; Schrader et al., 2012).

All PCR-based methods, DNA barcoding, metabarcoding, and mini-barcoding are based on the amplification and designing primers that amplify specifically only the selected region in target taxa genome/genomes. In addition to the long list of factors that may influence the success of PCR and have been studied by several investigators (the concentrations of the PCR buffer reagents, the primer length and the GC contents of primers and template, simple repeats in the primer sequence, stable secondary structures of both primer and product sequences, etc), careful design of primers is crucial for the success of PCR (Andreson et al., 2008). Oligonucleotide length, melting temperature, secondary structures, and primer dimer formation, as well as the specificity, should be considered for the in-silico selection of oligonucleotides.

The specificity of the oligonucleotides is one of the most important factors for good PCR. However, single-stranded DNA molecules also tend to bind to unintended targets or themselves. Optimal primers should hybridize only to the target sequence, particularly when complex genomic DNA is used as the template.

Ideally, there should be a pair of unique primers that amplify the desired target sequence selectively with maximum yield. This implies the prevention of unspecific binding to other DNA, such as non-target DNA, and other regions within a template (Hendling and Barišić, 2019).

One of the most important factors in determining PCR failure is the number of predicted primer-binding sites in the genomic DNA (Andreson et al., 2008). Alternative binding sites create unwanted amplifications, lower the yield of primers that can hybridize with the desired region or product, and therefore should be avoided whenever possible. Non-unique PCR primers can lead to amplification of undesired regions or no amplification, especially in experiments with complex genomes (Andreson et al., 2008).

Repeated DNA sequences are known to be significant components of genomes. Repetitive DNA sequences are abundant in a broad range of species, from bacteria to mammals. Genomes contain varying degrees of repetitive DNA (Haubold and Wiehe, 2006), and for many species, this represents the major fraction of the genome. Plant genomes contain particularly high proportions of repeats: for example, transposable elements cover >80% of the maize genome (Schnable et al., 2009). As in most plants, genes span smaller regions than in mammals, for instance, partly because large repetitive elements are rarely present inside plant introns. It has been estimated that >50% (Lander et al., 2001) and as much as 69% (de Koning et al., 2011) of the human genome is repetitive, however, over 80% of the genomes for some plant species is repetitive (Neale et al., 2014; Schnable et al., 2009).

The genomes of flowering plants vary in size from about 0.1 to over 100 giga-base pairs (Gbp), mostly because of polyploidy and variation in the abundance of repetitive elements in intergenic regions. Repetitive DNA constitutes a high proportion of plant genomes. Maize (*Zea mays*) is one of the most repetitive genomes for which a reference sequence is available (Schnable et al., 2009). Approximately 85% of the maize genome is composed of repetitive DNA (Hake and Walbot, 1980), mainly long terminal repeat (LTR)-retrotransposons (Rabinowicz and Bennetzen, 2006; SanMiguel et al., 1998, 1996). It seems that the multiplication of repeat sequences is the primary contributor to differences in DNA content between many (plant) taxa (Flavell et al., 1974). For example, barley and rice have similar complements of low-copy genes (Saghai Maroof et al., 1996) but a 12-fold difference in DNA content. The difference in DNA content is thus probably attributable to differences in the amount of repetitive DNA (Saghai Maroof et al., 1996). Most repeats in plant genomes are located in intergenic regions, but some are also located in coding sequences or pseudogenes (Gaut et al., 2000; Rabinowicz and Bennetzen, 2006; Sasaki, 2005). Repeats can also take the form of large-scale segmental duplications, such as those found on some human chromosomes (Zhang et al., 2005), and even whole-genome duplication, such as the duplication of the *Arabidopsis thaliana* genome (The Arabidopsis Genome Initiative, 2000).

The presence of a high amount of repetitive DNA regions in genomes has always presented technical challenges for sequence alignment and assembly programs (Alkan et al., 2011; Treangen and Salzberg, 2011). Repeats create ambiguities in alignment and assembly, which can produce biases and errors when interpreting results and cannot be ignored (Treangen and Salzberg, 2011). Repeated regions in genomes also pose a significant challenge to design primers that will hybridize and prime only on the target region. Nonspecific primers used for targeted amplification-based approaches may produce off-target products (Andreson et al., 2008; Francis et al., 2017; Hommelsheim et al., 2014; Miura et al., 2005).

There are several computational tools now available that facilitate genome-aware primer design (Andreson et al., 2006; Qu et al., 2012; Rotmistrovsky et al., 2004; Schuler, 1997), but obtaining specific amplification of targeted sequences, especially for genomes with large amounts of repetitive DNA, is still a difficult problem (Francis et al., 2017). In terms of selecting primers for DNA amplification, most tools assess specificity according to whether primer pairs (not individual primers) are predicted not to produce off-target amplicons within a specified size range (Francis et al., 2017). This criterion can result in the selection of primers that individually produce single-strand synthesis products at multiple off-target sites (Francis et al., 2017). For instance, a primer with many perfectly complementary off-target binding sites could be selected by tools that use this approach. Such primers can create byproducts that act as mega-primers in PCR, leading to background amplification (Hommelsheim et al., 2014). This issue is expected to be exacerbated when working with repetitive genomes. For example, it has been found that the repetitive nature of the genome can prevent enrichment and hybridization-based enrichment can be difficult (Fu et al., 2010) or unsuccessful (Ma and Axtell, 2013) in maize due to its repeat content. Therefore, evaluating the specificity of individual primers could improve the selection of target-specific primer pairs. Ideally, the repeated regions that cause off-target binding sites would be masked before primer design, to avoid the design of non-specific primers.

1.2.2. Whole genome sequencing based methods for metagenomic analysis

The main disadvantage of PCR-based detection methods is that they target only DNA from species to which the PCR primers bind efficiently. This also holds in principle for barcoding methods that rely on the PCR amplification and subsequent massively parallel next-generation sequencing of amplicons from variable genomic or cell organelle DNA regions. Barcoding methods have been shown to very efficiently identify taxa within environmental or food-derived metagenomic samples (Coghlan et al., 2012; Steven G Newmaster et al., 2013; Tillmar et al., 2013; Zhou et al., 2013), but require separate assays to address the different domains of life (Ripp et al., 2014).

However, metagenomic samples often contain DNA from very different organisms (e.g., plants, animals, bacteria, fungi, etc.) and different type of genomes (nuclear, mitochondrial, and in plants also plastid genomes). To overcome the limitations of amplification-based methods, a high-throughput sequencing method called next-generation sequencing (NGS) has been used.

Untargeted deep sequencing of total genomic DNA from the sample, followed by bioinformatic analysis of sequence reads, facilitates highly accurate identification of species from different kingdoms of life, and detection of sample components. In this case, the information used for the identification taxa is not restricted to only a few standard barcoding regions.

1.3. Identification of plants directly from metagenomic WGS data

In whole genome sequencing, extracted DNA is sequenced without pre-amplifying a specific DNA region through PCR. Sequence analysis of total genomic DNA isolated from food offers in principle the possibility to detect species in an unbiased way, enabling e.g., the detection of fraud in declared food composition or the identification of different allergens in food (Ripp et al., 2014).

In subsequent analysis, the DNA fragments (NGS reads) received from whole genome sequencing can be assembled using sophisticated software before the data can be analyzed further (Bayley, 2019). However, assembling the NGS (next-generation sequencing) reads offers challenges for computer memory, can be time-consuming, and prone to mistakes (Butler et al., 2008; Pop and Salzberg, 2008; Simpson and Durbin, 2012). Therefore, different assembly-free methods have been suggested for the identification of the taxonomical origin of DNA sequences from metagenomics sequencing. These methods are described in more detail below.

1.3.1. Alignment-based approach for WGS data analysis

One way to detect the taxonomic origin of sequencing reads can be by using the similarity of sequences. The most widely used alignment-based algorithm and method for the classification of the reads is the BLAST (the basic local alignment search tool) program (Altschul et al., 1990), which classifies a sequencing read by finding the best alignment to a database of genomic sequences. Different programs that use the BLAST algorithm have been published, e.g., MEGAN (Huson et al., 2007) and PhymmBL (Brady and Salzberg, 2011, 2009). Those developed programs try to improve BLAST's accuracy but perform at a speed lower than BLAST, which itself takes very substantial CPU (central processing unit) time to align and identify the millions of sequences generated by a typical Illumina sequencing run. Metagenomics data sets can be very large size today and the identification speed is critically important (Wood and Salzberg, 2014).

Ripp and colleagues showed that deep sequencing of total DNA derived from foodstuff material and using mapping/alignment of the sequencing reads to publicly available reference genome sequences can be readily used to identify species components (Ripp et al., 2014). However, this study also brings out that the biggest limitation of this alignment-based pipeline in terms of time and costs is set by the massive BLAST routines necessary for the metagenomic analysis (Ripp et al., 2014), however, also uncertainty and errors in finding the correct alignment have been discussed as a cause of errors in the downstream analysis (Löytynoja, 2012).

1.3.2. Alignment-free approach for WGS data analysis

Methods that use reasonable amounts of memory and in minimal time have been searched for the identification of different taxa from metagenomic samples. To solve some problems in bioinformatics *k*-mer counting strategy is suggested and used as an important step in many bioinformatics applications for analyzing metagenomic sequencing data.

DNA sequences can be considered as strings containing characters (nucleotides) A, T, G, C, or N (N denotes an undetermined character). The *k*-mers of a sequence (or a concatenation of many DNA sequences) represent all the possible subsequences of length *k*, and *k*-mer counting is determining the number of their occurrences in a sequence, where *k* is a positive integer (Marçais and Kingsford, 2011). Counting of *k*-mers is simple and straightforward, however, it becomes challenging when billions of sequences (reads) generated by next-generation sequencing (NGS) techniques must be processed using reasonable amounts of memory and in minimal time.

In recent years, a large number of *k*-mer counting programs have been designed, such as Jellyfish (Marçais and Kingsford, 2011), KMC3 (Kokot et al., 2017), Gerbil (Erbert et al., 2017), DSK (Rizk et al., 2013), etc. One possible naive strategy for *k*-mer counting is to use a dictionary (a hash table), with *k*-mers as keys and their counts as values. This *k*-mer counting approach is used in programs DSK (Rizk et al., 2013), Gerbil (Erbert et al., 2017), and Jellyfish (Marçais and Kingsford, 2011), for example. However, analyzing billions of input sequencing reads, the *k*-mer counting process is very slow and computer memory is often overwhelmed (Manekar and Sathe, 2018). The other widely used strategy, the sorting approach, works by sorting all *k*-mers extracted from each read. Thus, *k*-mer frequencies can be easily counted because, after sorting, repeating *k*-mers lay at adjacent positions in the sorted list. This strategy is applied in GenomeTester4 (Kaplinski et al., 2015), KMC1 (Deorowicz et al., 2013), KMC2 (Deorowicz et al., 2015), KMC3 (Kokot et al., 2017), and Turtle (Roy et al., 2014).

Approaches to *k*-mer counting proposed so far have mainly targeted memory-efficient and time-efficient solutions (Manekar and Sathe, 2018). Time, CPU, and memory are bounded (limited) resources, whereas the disk can be considered as a plentiful resource. As disks are always cheaper than memory, many researchers have focused on using the disk-based/external memory/out-of-core approach, as opposed to the in-memory/internal memory approach, to reduce memory usage. Disk-based approaches may additionally use hundreds of gigabytes of disk space for large datasets such as the human or plant genomes (Manekar and Sathe, 2018).

In our workgroup disk-based GenomeTester4 (GListMaker) (Kaplinski et al., 2015) has been developed that uses the sorting approach for *k*-mer counting. The *k*-mers gathered from the input file are first stored in temporary arrays, sorted, and then counted. Temporary arrays (tables) with counting results are then merged to produce the final *k*-mer count list. GenomeTester4 uses multiple threads to speed up *k*-mer counting. The disk-based approach has a much lower memory requirement than in-memory approaches and is designed to make it possible to count *k*-mers in large genomic datasets, such as a human or plant genome dataset, on commodity hardware (Manekar and Sathe, 2018). Memory usage can be greatly reduced using a disk because *k*-mers are processed in chunks and stored on a disk. GenomeTester4 programs are also applied in the development of *k*-mer-based methods described in Ref I, Ref. II and Ref. III of this thesis.

Counting the *k*-mers in a DNA sequence is an important step in many applications. Analysis of *k*-mers has been widely applied in many genomic analyses, including metagenomic analysis (Manekar and Sathe, 2018; Wood et al., 2019; Wood and Salzberg, 2014). However, there is a lack of tools that efficiently and accurately could identify plant taxa from metagenomics samples. There are many examples for applications of *k*-mer-based methods in the detection of bacterial taxa in metagenomic samples (Breitwieser et al., 2018; Ounit et al., 2015; Roosaare et al., 2017; Wood et al., 2019; Wood and Salzberg, 2014), but not successfully applied for plant identification, especially to identify plants which do not have sequenced whole genomes. The program Kraken (Wood and Salzberg, 2014) has shown to be a perspective program that uses *k*-mer based approach for the taxonomic classification of metagenomic sequencing reads (e.g., for the identification of bacteria). However, the sizes of plant genomes are substantially bigger compared to bacteria or viruses, and creating and using the reference database containing whole genome sequences of plants can be challenging in terms of memory usage. Kraken´s memory requirements can easily exceed 100 GB (Ye et al., 2019), especially when the reference database includes large eukaryotic genomes (Knutson et al., 2017; Meiser et al., 2017). Recently, an improved version of Kraken program, Kraken 2 (Wood et al., 2019), has been developed that has reduced its memory usage and increased the classification speed, allowing greater amounts of reference genomic data to be used. Another software Bracken ((Lu et al., 2017) has been used to analyze the classification results of Kraken programs to estimate species- or genus-level abundance of sequences. However, additional testing with plants and substantial upgrading of the reference database of Kraken programs, to cover more plant taxa and more genomic regions for different plants, is needed to apply Kraken and Bracken programs successfully for the identification of different plant taxa from metagenomic samples.

1.4. Advantages of using plastid genome regions in plant identification

Developments of sequencing technology and its increasingly low costs and high yields have also dramatically increased the number of available plastid genome sequences that could be used for the identification of plant species (Jiao et al., 2019).

Compared with the nuclear genome, the plastid genome is small in size, generally stable, mechanical breakdown resistant, circular form, has usually higher copy number in cell and has a higher interspecific and lower intraspecific sequence variation (Jiao et al., 2019; Kim et al., 2015; Li et al., 2015). Plastid genomes are haploid and non-recombining, so they act as a single locus (Nock et al., 2011). Another advantage is that the plastid genome has been found only in plants and some protists, therefore using plastid genomes for DNA analysis may help to bypass DNA contamination from organisms without chloroplasts/plastids (e.g., animals and fungi) (Dong et al., 2014).

Plastid genomes contain both highly conserved genes and more variable regions. The complete plastid genome size is usually ranging from 110 to 160 kbp, which is much longer than the length of commonly used DNA barcodes and provides sufficient variation to discriminate closely related plants. The whole plastid genome sequence has been proposed as a super-barcode for species-level plant identification (Bi et al., 2018; Dong et al., 2015; Kane and Cronk, 2008; Li et al., 2015; Nock et al., 2011; Parks et al., 2009). It has been also suggested, that using the complete plastid genome as a marker circumvents possible issues with gene deletion and low PCR efficiency, compared to targeting single short regions from specific genes as barcodes (Huang et al., 2015).

Conventional approaches to plastid genome sequencing commonly involve purification or PCR amplification of the plastid genome before sequencing (Cronn et al., 2008; Parks et al., 2009). However, these approaches are relatively time-consuming. Nock et al. (2011) demonstrated that massive parallel sequencing platforms have the capacity to sequence the plastid genome at over 100 times coverage in a single lane without purification. Despite representing a small fraction of total DNA sequence, 0.04% in rice, the concentration of plastid genome sequence reads is high relative to nuclear sequence in total DNA preparations (Nock et al., 2011). Non-purified (total) DNA extractions also include plastid DNA which is sequenced during massively parallel sequencing runs but is usually treated as contaminating sequence for many applications. However, these plastid genome sequences can be used for the identification of plants (Nock et al., 2011).

2. AIMS OF THE STUDY

The main aim of this study was to develop sequence analysis methods for the identification of plant taxa from degraded metagenomic samples.

For that:

1. To analyze the extent of potential PCR failure caused by repeated regions in plant genomes and to integrate *k*-mer based filters for masking these regions into primer-design software Primer3.
2. To develop a method for the identification of plant taxa-specific *k*-mers from plastid genomes and to test the method using *Solanum lycopersicum* (tomato) as an example.
3. To develop an alignment-free method for the detection of plant taxa directly from the whole genome sequencing reads of metagenomic samples and to test the method using lupin (*Lupinus* spp.) in food samples as a target.

3. RESULTS AND DISCUSSION

3.1. Identifying and masking PCR failure-prone regions in plant genomes (Ref. I)

3.1.1. Primer3_masker

Most of the methods for the identification of plants from metagenomics samples are based on the PCR amplification of a specific genomic region with target-specific oligonucleotides (including PCR, DNA barcoding, metabarcoding). PCR primer design may be complicated for eukaryotic genomes like plants that often contain a large number of repeat sequences and other regions that are unsuitable for amplification by PCR. Current methods are not efficient enough to avoid PCR primer design on the non-specific primer-binding sites.

Newly developed software "primer3_masker" (introduced in Ref. I) integrates *k*-mer-based masking and PCR primer design software Primer3. The masking bases on the genome-wide frequencies of the 16-mers and the 11-mers overlapping with the given position in their 3′end. Using a statistical model, this program masks failure-prone regions on the template DNA before primer design. Pre-generated *k*-mer lists for masking are available not only for model organisms but for sequences of 196 animal and plant genomes.

3.1.2. The extent of masking in different plants

The fraction of nucleotides that are masked in the sequence depends on the organism. Plant genomes may be very large and may contain more repeated regions than other organisms. We analyzed the masking extent in genomes of four widely known plant species: wheat (*Triticum aestivum*), maize (*Zea mays*), barley (*Hordeum vulgare*), rice (*Oryza sativa*). Our results showed that some of the plant genomes were more excessively masked than human genomes. Approximately 76% of nucleotides in wheat, 71% in maize genomes (compared to 44% of the human genome) were masked (using default values for masking with primer3_masker). Compared to the other available masking tool RepeatMasker, the extent of masking wheat and maize was significantly lower compared to the primer3_masker (44% vs 76% in wheat genome and 5% vs 71% in maize genome), which is probably caused by the fact that RepeatMasker uses the database of only known and annotated repeats for masking, but primer3_masker finds the repeated regions by brute force for each genome individually and is not dependent on the list of previously described repeats.

Our results also showed that using different failure rate cutoff values or the number of masked nucleotides, the results may vary. Decreasing the failure rate cutoff value or increasing the number of masked nucleotides, the reliability of designed primers increases but the fraction of the genome that can be used for primer design decreased.

This tool helps to find and mask PCR failure regions caused by repeats and mask these before primer design. This tool also has ready-to-use *k*-mer lists for about 45 plant species (https://primer3.ut.ee/lists.htm) and allows also custom *k*-mer list creation for different other species, it helps to design more specific primers for different PCR-based methods to identify plants or other organisms from different samples.

3.2. Identification plant taxa specific *k*-mers (Ref. II, Ref. III)

3.2.1. Method for selection of specific *k*-mers

Polymerase chain reaction based methods (e.g., PCR and different barcoding methods) commonly used for plant DNA identification from metagenomic samples are based on only a limited number of pre-amplified genomic regions (in the length of hundreds or thousands of nucleotides), which are often inapplicable due to DNA degradation, low amplification success or low species discriminative power of selected genomic regions. New developments in the field of the identification of plants in degraded metagenomic samples (including food, herbal products, gut content, environmental samples) are moving toward using a combination of many different genomic regions with reduced lengths (e.g., using mini-barcodes) or deep sequencing of total genomic DNA from samples with various taxonomical composition, followed by the identification of the taxonomical origin of sequencing reads. If we have to identify only a few species from the sample (e.g., one or a few allergens), the PCR-based method can be a very cost-effective and sensitive method to apply. However, if the aim is to identify tens or hundreds of different plant species from the sample, designing oligonucleotide primers can be challenging, time-consuming, and may not be cost-effective anymore. Whole genome sequencing of total DNA from metagenomic sample help avoid some of the problems associated with targeted PCR-base methods, however, assembly- and alignment-free methods for the analysis would help to cope with some computational challenges related to the increasing amount of available genomic data and by-pass some error-prone steps.

We developed *k*-mer based analysis method for the identification of a set of plant taxon-specific *k*-mers from the plastid genome (introduced in Ref. II and applied also in Ref. III). The identified *k*-mers can be used for the qualitative detection of plant taxa directly from raw sequencing reads, without assembling or aligning the reads. The pipeline converts assembled plastid genome sequences to *k*-mer lists and compares different lists of *k*-mers to find intersections, differences, or unions between the lists (Fig. 1 in Ref. II). Operations with *k*-mer lists are computationally effective and can be done fast (Kaplinski et al., 2015). The set of plant taxa-specific *k*-mers is identified using two steps for removing non-specific *k*-mers from the list of taxa-specific *k*-mers.

Plastid genome has been chosen for the identification of plant taxa specific *k*-mers mainly because of the higher copy number in cells compared to nuclear genome and higher number of available genome sequences for different plants in biological databases, however, using plastid genomes has also many other advantages.

3.2.2. *Solanum lycopersicum* (tomato) specific *k*-mers

We applied our method to identify *S. lycopersicum* (tomato or tomato plant) specific *k*-mers. *k*-mer length 32 nucleotides (nt) gave us the maximum number of species-specific *k*-mers (Fig. 4 in Ref. II). We identified 882 *S. lycopersicum* specific 32-mers that were present in at least two plastid genome sequences of *S. lycopersicum* and none of the 1714 plastid sequences from non-target species (Fig. 3 in Ref. II).

The identified 32-mers were located in 42 different regions from the *S. lycopersicum* plastid genome (Fig. 7 in Ref. II). All the *k*-mers were in single copy regions of the plastome, not in inverted repeats regions. This is in accordance with previous publications that have shown a lower mutation rate for regions of inverted repeats compared to single copy regions of the plastid genomes (Kahlau et al., 2006; Maier et al., 1995; Wolfe et al., 1987). Most of the previously described barcoding regions identified for plants are also mostly located in the single copy regions of the plastome.

Our results showed that *k*-mers identified from plastid genome sequences can be also detected from whole genome sequencing reads of *S. lycopersicum* (Fig. 5 in Ref. II). Probably due to the sequencing errors and sequence similarity of plastid genomes between phylogenetically close plant species, a small amount of *S. lycopersicum* *k*-mers can be also detected in whole genome sequencing reads from other *Solanum* species (e.g., *S. tuberosum* and *S. pimpinellifolium*), if the number of sequencing reads is high enough (Fig. 5 in Ref. II). However, it is possible to detect *S. lycopersicum* (tomato) in raw sequencing reads from metagenomic samples containing tomato, currant tomato, potato, eggplant, and bell pepper, using identified *S. lycopersicum* specific *k*-mers identified from the plastid genome, if approximately 600 (of 882) *S. lycopersicum* specific *k*-mers are detected (Fig. 5 in Ref. II). At least 100,000 sequencing reads from *S. lycopersicum* is needed to detect at least 600 *S. lycopersicum* specific *k*-mers.

Our results also showed that increased *k*-mers frequency cut-off values decrease the number of detected *S. lycopersicum* specific *k*-mers in the whole genome sequencing reads from nontarget species *S. tuberosum* and *S. pimpinellifolium*), though increases the minimal number of sequencing reads needed from *S. lycopersicum* to detect *S. lycopersicum* from the metagenomic sample), which may show that increased frequency cut-off value may increase the specificity, however with the cost of decreased sensitivity (Fig. 6 in Ref. II). However, additional testing with different samples is necessary to develop a mathematical method that correctly takes into consideration also the *k*-mers´ frequency value.

3.2.3. *Oryza sativa* and *Zea mays* specific *k*-mers

In addition to *S. lycopersicum* (tomato) specific *k*-mers, we also used our method to identify *Oryza sativa* (rice) and *Zea mays* (maize/corn) specific 32-mers and also analyzed the number of identified species-specific *k*-mers in whole genome sequencing reads from target and phylogenetically close nontarget species (Supplementary Fig. 2 in Ref. II).

The set of *O. sativa* specific *k*-mers contained 555 *k*-mers and the set of *Z. mays* specific *k*-mers contained 2304 *k*-mers (Supplementary Fig. 2 in Ref. II). The number of identified *k*-mers depends on different factors: the length of *k*-mer, the variability of target taxa sequences, the sequence similarity between target and non-target taxa, etc. Identifying taxa-specific *k*-mers for the identification of plant species that have widely used in plant breeding (e.g., wheat, rice), can be challenging and needs a very good and representative database of genomic sequences.

3.2.4. *Lupinus* specific *k*-mers

In Ref. III, we applied a developed pipeline also to identify *Lupinus* spp. (genus *Lupinus*), *Lupinus albus* (white lupin), *Lupinus luteus* (yellow lupin), and *Lupinus westianus* (gulf coast lupin) specific *k*-mers that could potentially be used to detect lupin as an allergen in food samples. We identified 31,179 genus-specific *k*-mers for *Lupinus* spp., 17,091 species-specific *k*-mers for *Lupinus albus*, 19,857 for *Lupinus luteus,* and 11,201 for *Lupinus westianus* from plastid genomes (Ref. III).

To assess the sensitivity of the method, we counted identified *Lupinus*-specific *k*-mers in the whole genome sequencing reads from the leaf or seed samples from different lupin species. We detected more than 30,000 of the 31,179 *Lupinus* spp. specific *k*-mers in the whole genome sequencing reads from species *L. albus, L. luteus,* and *L. westianus*, and more than 25,000 of the 31,179 *Lupinus* spp. *k*-mers in the species *L. angustifolius* and *L. mutabilis* (Fig. 1 in Ref. III). Plastid genome sequences for the two lupin species *L. angustifolius* and *L. mutabilis* were not available in sequence databases and were therefore not included in selecting a set of genus *Lupinus* specific *k*-mers. It shows that the composition and data representation in the sequence database is crucial to cover the variability of target taxa sequences and to provide sufficient universality (in target taxa) as well as specificity of the identified *k*-mers.

Our results with whole genome sequencing reads from phylogenetically close species showed that at least 1,500 of the 31,179 *Lupinus* spp. *k*-mers should be detected in the sequencing reads from the metagenomic sample that may contain also other leguminous species (like peanut or soy) to confirm the presence of lupin in the sample (Fig. 1 in Ref. III). At least 10,000 sequencing reads from lupin DNA are required to detect at least 1,500 *Lupinus* spp. specific *k*-mers. However, the cut-off value of the minimum number of required plant taxa-specific *k*-mers that are necessary to confirm the presence of plant taxa depends on the number of sequencing reads of the metagenomic sample. The cut-off value of 1,500 *k*-mers for lupin assumes the possibility that the yield of metagenomic sequencing data is at least 10^8 reads and almost all of these are from phylogenetically close species, however, as routine analysis of food or environmental samples this may be not cost-effective and much lower cut-off value may be used.

Analyzing species-specific *k*-mers, more than 90% of the 17,091 *L. albus*, 19,857 *L. luteus* or 11,201 *L. westianus* specific *k*-mers from plastid genome were detectable in the whole genome sequencing reads of *L. albus*, *L. luteus,* or *L. westianus,* respectively, if at least 250,000–500,000 reads are from lupin species *L. albus*, *L. luteus* or *L. westianus*, respectively (Fig. 2 in Ref. III and Supplementary Fig. 1 and 2 in Ref. III). 7,500 or more detected *Lupinus albus* specific *k*-mers, 4,300 or more *L. westianus* specific and 4,000 or more detected *L. luteus* specific *k*-mers confirms the presence of DNA from lupin species *L. albus, L. westianus,* and *L. luteus* in the metagenomic sample, even if the number of sequencing reads is 10^8 and contains predominantly of other leguminous species or other nontarget *Lupinus* species (Fig. 2 in Ref. III and Supplementary Fig. 1 and 2 in Ref. III).

3.3. Proof-of-principle study to identify plant DNA from processed food (Ref. III)

Food authentication is an important issue for the food industry to detect undeclared ingredients in food products (e.g., allergens) that may pose serious health risks to consumers. Using hundreds of taxon-specific short *k*-mers from all over the genome for the identification of plant taxa would give improved resolution at the species level detection as well as aid in analyzing complex degraded samples when sequence barcoding or other traditional PCR-based methods fail. Using whole genome sequencing and *k*-mer based method in the analysis of metagenomic sequencing reads in combination, enables to by-pass prior primer design and amplification of specific regions, genome assembly and mapping of sequencing reads to a reference genome.

To apply the identified plant taxa-specific *k*-mers to detect plants from real metagenomic samples (e.g., food or environmental samples), additional testing with real metagenomic samples is required. We used lupin as an example and applied our developed alignment-free method to identify short lupin-specific *k*-mers and to detect selected lupin-specific *k*-mers directly from whole genome sequencing reads from different samples.

3.3.1. Detection of lupin-specific *k*-mers in seed samples

The edible parts of plants are very often not green parts of the plant, but fruits or seeds that are assumed to contain a lower number of copies of plastid genomes in cells. We used *k*-mers from the plastid genome sequence for the identification plants and were interested if lupin-specific *k*-mers identified from the plastid genome are also detectable in whole-genome sequencing reads from non-green edible parts of the lupin plant (from seeds). Our results in Ref. III of this thesis showed that plant taxa-specific *k*-mers identified from the plastid genome are also detectable in whole-genome sequencing reads from seed. However, a small difference was observed between the samples of leaf and seed samples. We detected about 25,000–30,000 *Lupinus*-specific *k*-mers in the leaf or seedling sample if the number of whole genome sequencing reads was at least 100,000 (Fig. 1 in Ref. III). The same number of *k*-mers were detected in lupin seed samples if the number of whole genome sequencing reads was at least 500,000. It shows that *k*-mers identified from the plastid genome are identifiable also from the whole genome sequencing data from seeds; however, the sensitivity of detection may be slightly decreased.

3.3.2. Detection of lupin-specific *k*-mers in a processed food

The food matrix and processing may influence the detection of plants from metagenomic samples. Previous studies have shown a negative effect on the sensitivity of the method (Villa et al., 2018; Waiblinger et al., 2014). Our results with lupin showed that milling and short-term thermal processing do not alter substantially the detection of lupin with the *k*-mer based method. The lupin-specific *k*-mers were detected with similar sensitivity from sequencing reads of raw lupin seeds or of processed samples, like from the commercial lupin flour and canned (heated and salted) seeds (Fig. 1 in Ref. III).

Our proof-of-principle experiments with baked cookies (containing butter, sugar, salt, wheat flour, and different amount of lupin flour made from *L. angustifolius* seeds) showed that more than 1,500 lupin-specific *k*-mers (which is the minimum required amount of reads to detect lupin) were detected from sequencing reads from cookie sample that contains 0.05% or more lupin flour in flour mix (i.e. 0.02% or more lupin flour in a cookie), if the number of sequencing reads per sample was at least 19–35 million reads. If the lupin flour content in the flour mix is 5% or more the maximum amount of lupin-specific *k*-mers (about 25,000) were detected (Fig. 3 in Ref. III).

3.3.3. Testing the sensitivity of detection

Our results showed that the number of detected *k*-mers depends on the number of sequencing reads per sample and more sequencing reads were needed to detect the same number of *Lupinus* spp. specific *k*-mers in the cookie samples with lower lupin contents. Approximately 1–10 million sequencing reads from the food sample were sufficient to detect lupin flour content 0.5, 5, or 50% in wheat flour. However, at least 35 million reads were required to detect a lupin content of 0.05% in wheat flour (~0.02% lupin flour in the cookie), and even more, reads would be needed to detect a lupin content of 0.005% (Fig. 4 in Ref. III).

Our whole genome sequencing data analysis combined with the *k*-mers-based method using hundreds of short *k*-mers from different regions of the genome represents a good alternative to traditional amplification-based methods that use only one or a few amplifiable target genomic regions and often fail when analyzing the composition of complex and processed metagenomic samples containing degraded DNA (Carvalho et al., 2017; Lo and Shaw, 2018; Shokralla et al., 2015). The main limiting factor associated with whole genome sequencing based methods is often the high cost. However, the cost of the analysis is balanced by the abundant information derived from one whole genome sequencing run to answer different questions about the sample, e.g for food samples: to detect allergenic, toxic, or endangered species, to identify pathogenic bacteria, to detect fungi, viruses, etc. The *k*-mer based method can be easily multiplexed and used to simultaneously detect different species from the same metagenomic data-set using different *k*-mer sets for different target taxa or using different automated bioinformatic pipelines to answer different questions using the same sequencing data.

However, the application of the developed *k*-mer-based method in routine analyses to detect plant taxa from metagenomic samples like food or natural medicine products requires additional testing with different plant taxa and different types of samples.

CONCLUSIONS

More innovative, sensitive, and accurate analytical methods are needed to identify the composition of degraded metagenomics samples from different fields (food and herbal medicine products, environmental samples, etc.). The availability and decreased cost of next-generation sequencing, as well as the development of more effective algorithms for data analysis, have contributed to the development of new alternative methods and more effective pipelines for data analysis.

PCR-based methods are cost-effective and sensitive methods to detect one or a few target taxons from metagenomics samples. Designing primers that efficiently bind only targeted genomic regions can be challenging. New developments of Primer3 tools allow masking repeated regions of genomes (including plants) before primer design to prevent primers that lead to ineffective amplification or failure of the amplification reaction.

To overcome the limitations related to commonly used amplification-based methods and alignment-based data analysis approaches, whole genome sequencing based methods and more effective *k*-mer based pipelines for analyzing metagenomics data analysis have already been introduced for the identification of bacteria from metagenomics samples. The *k*-mer-based method for the whole genome sequencing data analysis introduced in the publications, related to this thesis, is a novel approach to identify plants from metagenomics samples.

We introduced the method to rapidly identify all short plant taxa-specific *k*-mers (maximum length of 32 nucleotides) from plastid genome sequences. These identified plant taxa-specific *k*-mers can be detected directly from sequencing reads from metagenomics samples to identify the presence of target plant taxa in the sample. Short *k*-mers from the plastid genome are detectable also from processed metagenomics samples containing degraded DNA (e.g from food). The sequencing-based method introduced combines next-generation sequencing with alignment- and assembling-free sequencing data analysis and represents a good alternative to the methods that are currently used to identify plants from different metagenomics samples.

SUMMARY IN ESTONIAN

Taimede DNA tuvastamine metagenoomsetest proovidest

Erinevate organismide DNA taksonoomiline tuvastamine paljusid erinevaid komponente sisaldavatest lagunenud või töödeldud keskkonnaproovidest (sh toiduainetest) võib olla keeruline väljakutse. Allergeenseid või muul põhjusel olulisi taimeliike saab edukalt toidust tuvastada DNA-põhiste meetoditega. Teise põlvkonna sekveneerimise kättesaadavus ja alanenud hind ning efektiivsemad andmeanalüüsi algoritmid võimaldavad töötada välja uusi alternatiivseid ja efektiivsemaid meetodeid metagenoomsete proovide taksonoomilise koosseisu kirjeldamiseks ning mahukate sekveneerimisandmete analüüsiks.

Metagenoomsete proovide analüüsiks laialdaselt kasutatavad PCR-põhised meetodid võimaldavad suhteliselt odavalt, tundlikult ja spetsiifiliselt detekteerida ühte või mõnda üksikut valitud liiki metagenoomsest proovist. Samas vajavad kõik amplifikatsiooni-põhised meetodid (sh PCR, DNA triipkoodi-meetodid) eelnevat PCR praimerite disaini, mis võib komplekssete, korduste-rikaste genoomijärjestuste puhul olla väga keerukas. Praimeridisaini programmi Primer3 tööriistapaketti lisandunud programm primer3_masker võimaldab juba enne praimerite disaini maskeerida regioonid, millele disainitud praimeritega amplifikatsioon suure tõenäosusega kordusjärjestustest tingitult ebaõnnestub või toimub ebaefektiivselt. Antud programm kasutab maskeerimiseks *k*-meeride põhist lähenemist ning võimaldab maskeerida ka taimegenoomide järjestusi, mille puhul kordusjärjestused võivad moodustada suure osa genoomist.

Kõik amplifikatsiooni-põhised meetodid (sh DNA triipkoodi-meetodid) põhinevad üksikute suhteliselt lühikeste genoomipiirkondade amplifikatsioonil, kasutades nii taksonite eristamiseks vaid piiratud osa genoomijärjestustes sisalduvast informatsioonist. Tihti kasutatakse järjestuste võrdlemisel joondamispõhiseid lähenemisi, mis võivad olla mahukamate genoomiandmete puhul ebaefektiivsed ja kulukad. Et vältida amplifikatsiooni-põhiste meetodite ja joondamispõhiste lähenemiste levinud puuduseid (ebaõnnestunud amplifikatsioonist tingitud valenegatiivsed ja valepositiivsed tulemused, joondamisvead, analüüsi kulukus jm), kasutatakse üha enam ülegenoomse sekveneerimise meetodeid ja *k*-meeridel põhinevaid lähenemisi metagenoomsete sekveneerimisandmete analüüsil. Efektiivseid *k*-meeride põhiseid lähenemisi on edukalt rakendatud juba bakterite tuvastamisel metagenoomsetest proovidest, kuid taimede tuvastamisel on antud doktoritööga seotud publikatsioonides kirjeldatud *k*-meeride-põhine tuvastamine veel uudne metoodika.

Oma publikatsioonides tutvustame metoodikat taimede taksonite spetsiifiliste (kuni 32 nukleotiidi pikkuste) *k*-meeride tuvastamiseks plastiidi genoomidest. Leitud *k*-meerid on tuvastatavad otse ülegenoomse sekveneerimise assambleerimata lugemitest ning neid saab kasutada taimede taksonite tuvastamiseks metagenoomsest proovist. Lühikesed *k*-meerid plastiidi genoomist on tuvastatavad ka lagunenud DNA-d sisaldavatest töödeldud metagenoomsetest proovidest (nt toidust). Väljatöötatud sekveneerimispõhine metoodika kombineerib teise põlvkonna sekvneerimise joondamis- ja assambleerimisvaba andmeanalüüsi metoodikaga pakkudes head alternatiivi meetoditele, mis on seni kasutusel olnud taimede tuvastamiseks metagenoomsetest proovidest.

REFERENCES

Ahmed, I., Biggs, P.J., Matthews, P.J., Collins, L.J., Hendy, M.D., Lockhart, P.J., 2012. Mutational Dynamics of Aroid Chloroplast Genomes. Genome Biol Evol 4, 1316–1323. https://doi.org/10.1093/gbe/evs110

Alkan, C., Coe, B.P., Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nature Reviews Genetics 12, 363–376. https://doi.org/10.1038/nrg2958

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andreson, R., Möls, T., Remm, M., 2008. Predicting failure rate of PCR in large genomes. Nucleic Acids Res 36, e66. https://doi.org/10.1093/nar/gkn290

Andreson, R., Reppo, E., Kaplinski, L., Remm, M., 2006. GENOMEMASKER package for designing unique genomic PCR primers. BMC Bioinformatics 7, 172. https://doi.org/10.1186/1471-2105-7-172

Arulandhu, A.J., Staats, M., Hagelaar, R., Voorhuijzen, M.M., Prins, T.W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F.B., Barreto Crespo, M.T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., Niekamp, H., Perri, E., Barbante, A., Rosec, J.-P., Seyfarth, R., Sovová, T., Van Moorleghem, C., van Ruth, S., Peelen, T., Kok, E., 2017. Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 6, 1–18. https://doi.org/10.1093/gigascience/gix080

Asahina, H., Shinozaki, J., Masuda, K., Morimitsu, Y., Satake, M., 2010. Identification of medicinal Dendrobium species by phylogenetic analyses using matK and rbcL sequences. J Nat Med 64, 133–138. https://doi.org/10.1007/s11418-009-0379-8

Baker, D.A., Stevenson, D.W., LittLe, D.P., 2012. DNA Barcode Identification of Black Cohosh Herbal Dietary Supplements. J AOAC Int 95, 1023–1034. https://doi.org/10.5740/jaoacint.11–261

Bauer, T., Weller, P., Hammes, W.P., Hertel, C., 2003. The effect of processing parameters on DNA degradation in food. Eur Food Res Technol 217, 338–343. https://doi.org/10.1007/s00217-003-0743-y

Bayley, A., 2019. A Summary of Current DNA Methods for Herb and Spice Identification. J AOAC Int 102, 386–389. https://doi.org/10.5740/jaoacint.18–0388

Bi, Y., Zhang, M., Xue, J., Dong, R., Du, Y., Zhang, X., 2018. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on Fritillaria. Sci Rep 8, 1184. https://doi.org/10.1038/s41598-018-19591-9

Bista, I., Carvalho, G.R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., Creer, S., 2018. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources 18, 1020–1034. https://doi.org/10.1111/1755-0998.12888

Bleeker, W., Klausmeyer, S., Peintinger, M., Dienst, M., 2008. DNA sequences identify invasive alien Cardamine at Lake Constance. Biological Conservation 141, 692–698. https://doi.org/10.1016/j.biocon.2007.12.015

Bosmali, I., Ganopoulos, I., Madesis, P., Tsaftaris, A., 2012. Microsatellite and DNA-barcode regions typing combined with High Resolution Melting (HRM) analysis for food forensic uses: A case study on lentils (Lens culinaris). Food Research International 46, 141–147. https://doi.org/10.1016/j.foodres.2011.12.013

Brady, A., Salzberg, S., 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8, 367. https://doi.org/10.1038/nmeth0511-367

Brady, A., Salzberg, S.L., 2009. Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models. Nat Methods 6, 673–676. https://doi.org/10.1038/nmeth.1358

Breitwieser, F.P., Baker, D.N., Salzberg, S.L., 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology 19, 198. https://doi.org/10.1186/s13059-018-1568-0

Bruni, I., Galimberti, A., Caridi, L., Scaccabarozzi, D., De Mattia, F., Casiraghi, M., Labra, M., 2015. A DNA barcoding approach to identify plant species in multiflower honey. Food Chemistry 170, 308–315. https://doi.org/10.1016/j.foodchem.2014.08.060

Bruno, A., Sandionigi, A., Agostinetto, G., Bernabovi, L., Frigerio, J., Casiraghi, M., Labra, M., 2019. Food Tracking Perspective: DNA Metabarcoding to Identify Plant Composition in Complex and Processed Food Products. Genes (Basel) 10, 248. https://doi.org/10.3390/genes10030248

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. 18, 810–820. https://doi.org/10.1101/gr.7337908

Carvalho, D.C., Palhares, R.M., Drummond, M.G., Gadanho, M., 2017. Food metagenomics: Next generation sequencing identifies species mixtures and mislabeling within highly processed cod products. Food Control 80, 183–186. https://doi.org/10.1016/j.foodcont.2017.04.049

CBOL Plant Working Group, 2009. A DNA barcode for land plants. Proc Natl Acad Sci U S A 106, 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chan, K., 2003. Some aspects of toxic contaminants in herbal medicines. Chemosphere 52, 1361–1371. https://doi.org/10.1016/S0045-6535(03)00471-5

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., Berg, C. van den, Madriñán, S., Petersen, G., Seberg, O., Jørgsensen, T., Cameron, K.M., Carine, M., Pedersen, N., Hedderson, T.A.J., Conrad, F., Salazar, G.A., Richardson, J.E., Hollingsworth, M.L., Barraclough, T.G., Kelly, L., Wilkinson, M., 2007. A proposal for a standardised protocol to barcode all land plants. TAXON 56, 295–299. https://doi.org/10.1002/tax.562004

Chen, F., Chan, H.-Y.E., Wong, K.-L., Wang, J., Yu, M.-T., But, P.P.-H., Shaw, P.-C., 2008. Authentication of Saussurea lappa, an Endangered Medicinal Material, by ITS DNA and 5S rRNA Sequencing. Planta Med 74, 889–892. https://doi.org/10.1055/s-2008-1074551

Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J., Leon, C., 2014. A renaissance in herbal medicine identification: From morphology to DNA. Biotechnology Advances 32, 1237–1244. https://doi.org/10.1016/j.biotechadv.2014.07.004

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., Leon, C., 2010. Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. PLoS One 5, e8613. https://doi.org/10.1371/journal.pone.0008613

Cheng, X., Su, X., Chen, X., Zhao, H., Bo, C., Xu, J., Bai, H., Ning, K., 2014. Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. Scientific Reports 4, 5147. https://doi.org/10.1038/srep05147

Cline, E., 2012. Marketplace substitution of Atlantic salmon for Pacific salmon in Washington State detected by DNA barcoding. Food Research International 45, 388–393. https://doi.org/10.1016/j.foodres.2011.10.043

Coghlan, M.L., Haile, J., Houston, J., Murray, D.C., White, N.E., Moolhuijzen, P., Bellgard, M.I., Bunce, M., 2012. Deep Sequencing of Plant and Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health Safety Concerns. PLOS Genetics 8, e1002657. https://doi.org/10.1371/journal.pgen.1002657

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R., Mockler, T., 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36, e122–e122. https://doi.org/10.1093/nar/gkn502

Danezis, G.P., Tsagkaris, A.S., Camin, F., Brusic, V., Georgiou, C.A., 2016. Food authentication: Techniques, trends & emerging approaches. TrAC Trends in Analytical Chemistry, On-site and In-vivo Instrumentation and Applications 85, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

de Boer, H.J., Ichim, M.C., Newmaster, S.G., 2015. DNA Barcoding and Pharmacovigilance of Herbal Medicines. Drug Saf 38, 611–620. https://doi.org/10.1007/s40264-015-0306-8

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genet 7. https://doi.org/10.1371/journal.pgen.1002384

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol Lett 10, 20140562. https://doi.org/10.1098/rsbl.2014.0562

Deorowicz, S., Debudaj-Grabysz, A., Grabowski, S., 2013. Disk-based k-mer counting on a PC. BMC Bioinformatics 14, 160. https://doi.org/10.1186/1471-2105-14-160

Deorowicz, S., Kokot, M., Grabowski, S., Debudaj-Grabysz, A., 2015. KMC 2: fast and resource-frugal k-mer counting. Bioinformatics 31, 1569–1576. https://doi.org/10.1093/bioinformatics/btv022

Di Pinto, A., Mottola, A., Marchetti, P., Bottaro, M., Terio, V., Bozzo, G., Bonerba, E., Ceci, E., Tantillo, G., 2016. Packaged frozen fishery products: species identification, mislabeling occurrence and legislative implications. Food Chemistry 194, 279–283. https://doi.org/10.1016/j.foodchem.2015.07.135

Dong, W., Liu, H., Xu, C., Zuo, Y., Chen, Z., Zhou, S., 2014. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: A case study on ginsengs. ResearchGate 15, 138. https://doi.org/10.1186/s12863-014-0138-z

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., Zhou, S., 2015. ycf1, the most promising plastid DNA barcode of land plants. Sci Rep 5, 8348. https://doi.org/10.1038/srep08348

Egan, A.N., Schlueter, J., Spooner, D.M., 2012. Applications of next-generation sequencing in plant biology. American Journal of Botany 99, 175–185. https://doi.org/10.3732/ajb.1200020

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15. https://doi.org/10.1038/hdy.2010.152

Elbrecht, V., Leese, F., 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass – Sequence Relationships with an Innovative Metabarcoding Protocol. PLOS ONE 10, e0130324. https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., Vamos, E.E., Steinke, D., Leese, F., 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6, e4644. https://doi.org/10.7717/peerj.4644

Erbert, M., Rechner, S., Müller-Hannemann, M., 2017. Gerbil: a fast and memory-efficient k-mer counter with GPU-support. Algorithms Mol Biol 12, 9. https://doi.org/10.1186/s13015-017-0097-9

Eurlings, M.C.M., Lens, F., Pakusza, C., Peelen, T., Wieringa, J.J., Gravendeel, B., 2013. Forensic Identification of Indian Snakeroot (Rauvolfia serpentina Benth. ex Kurz) Using DNA Barcoding. Journal of Forensic Sciences 58, 822–830. https://doi.org/10.1111/1556-4029.12072

Fahner, N.A., Shokralla, S., Baird, D.J., Hajibabaei, M., 2016. Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. PLOS ONE 11, e0157505. https://doi.org/10.1371/journal.pone.0157505

Fang, X., Zhang, C., 2016. Detection of adulterated murine components in meat products by TaqMan© real-time PCR. Food Chemistry 192, 485–490. https://doi.org/10.1016/j.foodchem.2015.07.020

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M., Barrett, S.C.H., 2008. Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLoS One 3. https://doi.org/10.1371/journal.pone.0002802

Ferri, G., Alù, M., Corradini, B., Beduschi, G., 2009. Forensic botany: species identification of botanical trace evidence using a multigene barcoding approach. Int J Legal Med 123, 395–401. https://doi.org/10.1007/s00414-009-0356-5

Flavell, R.B., Bennett, M.D., Smith, J.B., Smith, D.B., 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12, 257–269. https://doi.org/10.1007/BF00485947

Francis, F., Dumas, M.D., Wisser, R.J., 2017. ThermoAlign: a genome-aware primer design tool for tiled amplicon resequencing. Sci Rep 7, 44437. https://doi.org/10.1038/srep44437

Fu, Y., Springer, N.M., Gerhardt, D.J., Ying, K., Yeh, C.-T., Wu, W., Swanson-Wagner, R., D'Ascenzo, M., Millard, T., Freeberg, L., Aoyama, N., Kitzman, J., Burgess, D., Richmond, T., Albert, T.J., Barbazuk, W.B., Jeddeloh, J.A., Schnable, P.S., 2010. Repeat subtraction-mediated sequence capture from a complex genome. Plant J 62, 898–909. https://doi.org/10.1111/j.1365-313X.2010.04196.x

Galimberti, A., Sandionigi, A., Bruno, A., Bruni, I., Barbuto, M., Casiraghi, M., Labra, M., 2015. Towards a Universal Molecular Approach for the Quality Control of New Foodstuffs, in: Advances in Food Biotechnology. John Wiley & Sons, Ltd, pp. 37–60. https://doi.org/10.1002/9781118864463.ch04

Gao, Z., Liu, Y., Wang, X., Wei, X., Han, J., 2019. DNA Mini-Barcoding: A Derived Barcoding Method for Herbal Molecular Identification. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.00987

Gathier, G., Niet, T. van der, Peelen, T., Vugt, R.R. van, Eurlings, M.C.M., Gravendeel, B., 2013. Forensic Identification of CITES Protected Slimming Cactus (Hoodia) Using DNA Barcoding. Journal of Forensic Sciences 58, 1467–1471. https://doi.org/10.1111/1556-4029.12184

Gaut, B.S., d'Ennequin, M.L.T., Peek, A.S., Sawkins, M.C., 2000. Maize as a model for the evolution of plant nuclear genomes. PNAS 97, 7008–7015. https://doi.org/10.1073/pnas.97.13.7008

Ghorbani, A., Saeedi, Y., de Boer, H.J., 2017. Unidentifiable by morphology: DNA barcoding of plant material in local markets in Iran. PLoS One 12, e0175722. https://doi.org/10.1371/journal.pone.0175722

Gryson, N., Ronsse, F., Messens, K., De Loose, M., Verleyen, T., Dewettinck, K., 2002. Detection of DNA during the refining of soybean oil. J Amer Oil Chem Soc 79, 171–174. https://doi.org/10.1007/s11746-002-0453-2

Hajibabaei, M., Singer, G.A.C., Hebert, P.D.N., Hickey, D.A., 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends in Genetics 23, 167–172. https://doi.org/10.1016/j.tig.2007.02.001

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B., Hebert, P.D.N., 2006. A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes 6, 959–964. https://doi.org/10.1111/j.1471-8286.2006.01470.x

Hake, S., Walbot, V., 1980. The genome of Zea mays, its organization and homology to related grasses. Chromosoma 79, 251–270. https://doi.org/10.1007/BF00327318

Haubold, B., Wiehe, T., 2006. How repetitive are genomes? BMC Bioinformatics 7, 541. https://doi.org/10.1186/1471-2105-7-541

Hebert, P.D.N., Gregory, T.R., 2005. The Promise of DNA Barcoding for Taxonomy. Syst Biol 54, 852–859. https://doi.org/10.1080/10635150500354886

Hebert, P.D.N., Ratnasingham, S., deWaard, J.R., 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc Biol Sci 270, S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Hendling, M., Barišić, I., 2019. In-silico Design of DNA Oligonucleotides: Challenges and Approaches. Computational and Structural Biotechnology Journal 17, 1056–1065. https://doi.org/10.1016/j.csbj.2019.07.008

Heubl, G., 2010. New Aspects of DNA-based Authentication of Chinese Medicinal Plants by Molecular Biological Techniques. Planta Med 76, 1963–1974. https://doi.org/10.1055/s-0030-1250519

Hollingsworth, P.M., 2008. DNA barcoding plants in biodiversity hot spots: Progress and outstanding questions. Heredity 101, 1–2. https://doi.org/10.1038/hdy.2008.16

Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and Using a Plant DNA Barcode. PLoS One 6, e19254. https://doi.org/10.1371/journal.pone.0019254

Hommelsheim, C.M., Frantzeskakis, L., Huang, M., Ülker, B., 2014. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep 4, 5052. https://doi.org/10.1038/srep05052

Hossain, M.A.M., Ali, Md.E., Sultana, S., Asing, Bonny, S.Q., Kader, Md.A., Rahman, M.A., 2017. Quantitative Tetraplex Real-Time Polymerase Chain Reaction Assay with TaqMan Probes Discriminates Cattle, Buffalo, and Porcine Materials in Food Chain. J. Agric. Food Chem. 65, 3975–3985. https://doi.org/10.1021/acs.jafc.7b00730

Hrnčírová, Z., Bergerová, E., Siekel, P., 2008. Effects of technological treatment on DNA degradation in selected food matrices of plant origin. Journal of Food and Nutrition Research (Slovak Republic) 47, 23–28.

Huang, W., Li, F., Liu, Y., Long, C., 2015. Identification of Crocus sativus and its Adulterants from Chinese Markets by using DNA Barcoding Technique. Iran J Biotechnol 13, 36–42. https://doi.org/10.15171/ijb.1034

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res 17, 377–386. https://doi.org/10.1101/gr.5969107

Jaakola, L., Suokas, M., Häggman, H., 2010. Novel approaches based on DNA barcoding and high-resolution melting of amplicons for authenticity analyses of berry species. Food Chemistry 123, 494–500. https://doi.org/10.1016/j.foodchem.2010.04.069

Janjua, S., Fakhar-I-Abbas, null, William, K., Malik, I.U., Mehr, J., 2017. DNA Mini-barcoding for wildlife trade control: a case study on identification of highly processed animal materials. Mitochondrial DNA A DNA Mapp Seq Anal 28, 544–546. https://doi.org/10.3109/24701394.2016.1155051

Jiao, L., Lu, Y., He, T., Li, J., Yin, Y., 2019. A strategy for developing high-resolution DNA barcodes for species discrimination of wood specimens using the complete chloroplast genome of three Pterocarpus species. Planta 250, 95–104. https://doi.org/10.1007/s00425-019-03150-1

Jiao, L., Yu, M., Wiedenhoeft, A.C., He, T., Li, J., Liu, B., Jiang, X., Yin, Y., 2018. DNA Barcode Authentication and Library Development for the Wood of Six Commercial Pterocarpus Species: the Critical Role of Xylarium Specimens. Sci Rep 8, 1945. https://doi.org/10.1038/s41598-018-20381-6

Kahlau, S., Aspinall, S., Gray, J.C., Bock, R., 2006. Sequence of the Tomato Chloroplast DNA and Evolutionary Comparison of Solanaceous Plastid Genomes. J Mol Evol 63, 194–207. https://doi.org/10.1007/s00239-005-0254-5

Kane, N.C., Cronk, Q., 2008. Botany without borders: barcoding in focus. Molecular Ecology 17, 5175–5176. https://doi.org/10.1111/j.1365-294X.2008.03972.x

Kaplinski, L., Lepamets, M., Remm, M., 2015. GenomeTester4: a toolkit for performing basic set operations – union, intersection and complement on k-mer lists. Gigascience 4, 58. https://doi.org/10.1186/s13742-015-0097-y

Khan, I.A., Smillie, T., 2012. Implementing a "Quality by Design" Approach to Assure the Safety and Integrity of Botanical Dietary Supplements. J. Nat. Prod. 75, 1665–1673. https://doi.org/10.1021/np300434j

Kim, K., Lee, S.-C., Lee, J., Lee, H.O., Joh, H.J., Kim, N.-H., Park, H.-S., Yang, T.-J., 2015. Comprehensive Survey of Genetic Diversity in Chloroplast Genomes and 45S nrDNAs within Panax ginseng Species. PLOS ONE 10, e0117159. https://doi.org/10.1371/journal.pone.0117159

Knutson, T.P., Velayudhan, B.T., Marthaler, D.G., 2017. A porcine enterovirus G associated with enteric disease contains a novel papain-like cysteine protease. Journal of General Virology, 98, 1305–1310. https://doi.org/10.1099/jgv.0.000799

Kokot, M., Długosz, M., Deorowicz, S., 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33, 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kress, W.J., Erickson, D.L., 2007. A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. PLOS ONE 2, e508. https://doi.org/10.1371/journal.pone.0000508

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, Christina, Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, Christopher, Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, C. for G.R., The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Department of Genome Analysis, I. of M.B., GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, Multimegabase Sequencing Center, T.I. for S.B., Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Cold Spring Harbor Laboratory, L.A.H.G.C., GBF – German Research Centre for Biotechnology:, *Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):, Scientific management: National Human Genome Research Institute, U.N.I. of H., Stanford Human Genome Center:, University of Washington Genome Center:, Department of Molecular Biology, K.U.S. of M., University of Texas Southwestern Medical Center at Dallas:, Office of Science, U.D. of E., The Wellcome Trust:, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Lee, S.Y., Ng, W.L., Mahat, M.N., Nazre, M., Mohamed, R., 2016. DNA Barcoding of the Endangered Aquilaria (Thymelaeaceae) and Its Application in Species Authentication of Agarwood Products Traded in the Market. PLoS One 11, e0154631. https://doi.org/10.1371/journal.pone.0154631

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T., Machida, R.J., 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front Zool 10, 34. https://doi.org/10.1186/1742-9994-10-34

Li, X., Yang, Y., Henry, R.J., Rossetto, M., Wang, Y., Chen, S., 2015. Plant DNA barcoding: from gene to genome. Biological Reviews 90, 157–166. https://doi.org/10.1111/brv.12104

Little, D.P., 2014a. Authentication of Ginkgo biloba herbal dietary supplements using DNA barcoding. Genome 57, 513–516. https://doi.org/10.1139/gen-2014-0130

Little, D.P., 2014b. A DNA mini-barcode for land plants. Molecular Ecology Resources 14, 437–446. https://doi.org/10.1111/1755-0998.12194

Little, D.P., Jeanson, M.L., 2013. DNA Barcode Authentication of Saw Palmetto Herbal Dietary Supplements. Sci Rep 3, 3518. https://doi.org/10.1038/srep03518

Lo, Y.-T., Shaw, P.-C., 2018. DNA-based techniques for authentication of processed food and food supplements. Food Chemistry 240, 767–774. https://doi.org/10.1016/j.foodchem.2017.08.022

Löytynoja, A., 2012. Alignment Methods: Strategies, Challenges, Benchmarking, and Comparative Overview, in: Anisimova, M. (Ed.), Evolutionary Genomics: Statistical and Computational Methods, Volume 1, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 203–235. https://doi.org/10.1007/978-1-61779-582-4_7

Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3, e104. https://doi.org/10.7717/peerj-cs.104

Ma, Z., Axtell, M.J., 2013. Long-Range Genomic Enrichment, Sequencing, and Assembly to Determine Unknown Sequences Flanking a Known microRNA. PLoS One 8, e83721. https://doi.org/10.1371/journal.pone.0083721

Madesis, P., Ganopoulos, I., Anagnostis, A., Tsaftaris, A., 2012. The application of Bar-HRM (Barcode DNA-High Resolution Melting) analysis for authenticity testing and quantitative detection of bean crops (Leguminosae) without prior DNA purification. Food Control 25, 576–582. https://doi.org/10.1016/j.foodcont.2011.11.034

Madesis, P., Ganopoulos, I., Sakaridis, I., Argiriou, A., Tsaftaris, A., 2014. Advances of DNA-based methods for tracing the botanical origin of food products. Food Research International, Authenticity, Typicality, Traceability and Intrinsic Quality of Food Products 60, 163–172. https://doi.org/10.1016/j.foodres.2013.10.042

Maier, R.M., Neckermann, K., Igloi, G.L., Kössel, H., 1995. Complete Sequence of the Maize Chloroplast Genome: Gene Content, Hotspots of Divergence and Fine Tuning of Genetic Information by Transcript Editing. Journal of Molecular Biology 251, 614–628. https://doi.org/10.1006/jmbi.1995.0460

Manekar, S.C., Sathe, S.R., 2018. A benchmark study of k-mer counting methods for high-throughput sequencing. Gigascience 7, giy125. https://doi.org/10.1093/gigascience/giy125

Marçais, G., Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770. https://doi.org/10.1093/bioinformatics/btr011

Meiser, A., Otte, J., Schmitt, I., Grande, F.D., 2017. Sequencing genomes from mixed DNA samples – evaluating the metagenome skimming approach in lichenized fungi. Scientific Reports 7, 14881. https://doi.org/10.1038/s41598-017-14576-6

Meusnier, I., Singer, G.A., Landry, J.-F., Hickey, D.A., Hebert, P.D., Hajibabaei, M., 2008. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9, 214. https://doi.org/10.1186/1471-2164-9-214

Miller Coyle, H., Lee, C.-L., Lin, W.-Y., Lee, H.C., Palmbach, T.M., 2005. Forensic botany: using plant evidence to aid in forensic death investigation. Croat Med J 46, 606–612.

Miura, F., Uematsu, C., Sakaki, Y., Ito, T., 2005. A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3′-end subsequences. Bioinformatics 21, 4363–4370. https://doi.org/10.1093/bioinformatics/bti716

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., Martínez-García, P.J., Vasquez-Gross, H.A., Lin, B.Y., Zieve, J.J., Dougherty, W.M., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marçais, G., Roberts, M., Holt, C., Yandell, M., Davis, J.M., Smith, K.E., Dean, J.F., Lorenz, W.W., Whetten, R.W., Sederoff, R., Wheeler, N., McGuire, P.E., Main, D., Loopstra, C.A., Mockaitis, K., deJong, P.J., Yorke, J.A., Salzberg, S.L., Langley, C.H., 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15, R59. https://doi.org/10.1186/gb-2014-15-3-r59

Newmaster, S. g., Fazekas, A. j., Ragupathy, S., 2006. DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. Can. J. Bot. 84, 335–341. https://doi.org/10.1139/b06-047

Newmaster, Steven G., Grguric, M., Shanmughanandhan, D., Ramalingam, S., Ragupathy, S., 2013. DNA barcoding detects contamination and substitution in North American herbal products. BMC Medicine 11, 222. https://doi.org/10.1186/1741-7015-11-222

Nock, C.J., Waters, D.L.E., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M., Henry, R.J., 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal 9, 328–333. https://doi.org/10.1111/j.1467-7652.2010.00558.x

Novak, J., Grausgruber-Gröger, S., Lukas, B., 2007. DNA-based authentication of plant extracts. Food Research International 40, 388–392. https://doi.org/10.1016/j.foodres.2006.10.015

Ouarghidi, A., Powell, B., Martin, G.J., De Boer, H., Abbad, A., 2012. Species Substitution in Medicinal Roots and Possible Implications for Toxicity of Herbal Remedies in Morocco. Econ Bot 66, 370–382. https://doi.org/10.1007/s12231-012-9215-2

Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16, 236. https://doi.org/10.1186/s12864-015-1419-2

Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol 7, 84. https://doi.org/10.1186/1741-7007-7-84

Parveen, I., Gafner, S., Techen, N., Murch, S.J., Khan, I.A., 2016. DNA Barcoding for the Identification of Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations. Planta Med 82, 1225–1235. https://doi.org/10.1055/s-0042-111208

Pawar, R.S., Handy, S.M., Cheng, R., Shyong, N., Grundel, E., 2017. Assessment of the Authenticity of Herbal Dietary Supplements: Comparison of Chemical and DNA Barcoding Methods. Planta Med 83, 921–936. https://doi.org/10.1055/s-0043-107881

Piñol, J., Senar, M.A., Symondson, W.O.C., 2019. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology 28, 407–419. https://doi.org/10.1111/mec.14776

Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet 24, 142–149. https://doi.org/10.1016/j.tig.2007.12.007

Prosser, S.W.J., Hebert, P.D.N., 2017. Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. Food Chemistry 214, 183–191. https://doi.org/10.1016/j.foodchem.2016.07.077

Qu, W., Zhou, Y., Zhang, Y., Lu, Y., Wang, X., Zhao, D., Yang, Y., Zhang, C., 2012. MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. Nucleic Acids Res 40, W205–W208. https://doi.org/10.1093/nar/gks552

Rabinowicz, P.D., Bennetzen, J.L., 2006. The maize genome as a model for efficient sequence analysis of large plant genomes. Current Opinion in Plant Biology, Genome studies and molecular genetics: Part 1: Model legumes / edited by Nevin D Young and Randy C Shoemaker; Part 2: Maize genomics / edited by Susan R Wessler. Plant biotechnology / edited by John Salmeron and Luis R Herrera-Estrella 9, 149–156. https://doi.org/10.1016/j.pbi.2006.01.015

Raclariu, A.C., Heinrich, M., Ichim, M.C., de Boer, H., 2018. Benefits and Limitations of DNA Barcoding and Metabarcoding in Herbal Product Authentication. Phytochem Anal 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu, A.C., Mocan, A., Popa, M.O., Vlase, L., Ichim, M.C., Crisan, G., Brysting, A.K., de Boer, H., 2017a. Veronica officinalis Product Authentication Using DNA Metabarcoding and HPLC-MS Reveals Widespread Adulteration with Veronica chamaedrys. Front Pharmacol 8, 378. https://doi.org/10.3389/fphar.2017.00378

Raclariu, A.C., Paltinean, R., Vlase, L., Labarre, A., Manzanilla, V., Ichim, M.C., Crisan, G., Brysting, A.K., de Boer, H., 2017b. Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci Rep 7, 1291. https://doi.org/10.1038/s41598-017-01389-w

Ripp, F., Krombholz, C.F., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., Hankeln, T., 2014. All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics 15, 639. https://doi.org/10.1186/1471-2164-15-639

Rizk, G., Lavenier, D., Chikhi, R., 2013. DSK: k-mer counting with very low memory usage. Bioinformatics 29, 652–653. https://doi.org/10.1093/bioinformatics/btt020

Röder, M., Vieths, S., Holzhauser, T., 2011. Sensitive and specific detection of potentially allergenic almond (Prunus dulcis) in complex food matrices by Taqman® real-time polymerase chain reaction in comparison to commercially available protein-based enzyme-linked immunosorbent assay. Analytica Chimica Acta 685, 74–83. https://doi.org/10.1016/j.aca.2010.11.019

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

Rotmistrovsky, K., Jang, W., Schuler, G.D., 2004. A web server for performing electronic PCR. Nucleic Acids Res 32, W108–W112. https://doi.org/10.1093/nar/gkh450

Roy, R.S., Bhattacharya, D., Schliep, A., 2014. Turtle: identifying frequent k-mers with cache-efficient algorithms. Bioinformatics 30, 1950–1957. https://doi.org/10.1093/bioinformatics/btu132

Saghai Maroof, M.A., Yang, G.P., Biyashev, R.M., Maughan, P.J., Zhang, Q., 1996. Analysis of the barley and rice genomes by comparative RFLP linkage mapping. Theoret. Appl. Genetics 92, 541–551. https://doi.org/10.1007/BF00224556

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L., 1998. The paleontology of intergene retrotransposons of maize. Nature Genetics 20, 43–45. https://doi.org/10.1038/1695

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768. https://doi.org/10.1126/science.274.5288.765

Särkinen, T., Staats, M., Richardson, J.E., Cowan, R.S., Bakker, F.T., 2012. How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens. PLOS ONE 7, e43808. https://doi.org/10.1371/journal.pone.0043808

Sasaki, T., 2005. The map-based sequence of the rice genome. Nature 436, 793–800. https://doi.org/10.1038/nature03895

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Buren, P.V., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., Wilson, R.K., 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326, 1112–1115. https://doi.org/10.1126/science.1178534

Schrader, C., Schielke, A., Ellerbroek, L., Johne, R., 2012. PCR inhibitors – occurrence, properties and removal. Journal of Applied Microbiology 113, 1014–1026. https://doi.org/10.1111/j.1365-2672.2012.05384.x

Schuler, G.D., 1997. Sequence Mapping by Electronic PCR. Genome Res 7, 541–550. https://doi.org/10.1101/gr.7.5.541

Seethapathy, G.S., Raclariu-Manolica, A.-C., Anmarkrud, J.A., Wangensteen, H., de Boer, H.J., 2019. DNA Metabarcoding Authentication of Ayurvedic Herbal Products on the European Market Raises Concerns of Quality and Fidelity. Front. Plant Sci. 10, 68. https://doi.org/10.3389/fpls.2019.00068

Shokralla, S., Hellberg, R.S., Handy, S.M., King, I., Hajibabaei, M., 2015. A DNA Mini-Barcoding System for Authentication of Processed Fish Products. Scientific Reports 5, 15894. https://doi.org/10.1038/srep15894

Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Simpson, J.T., Durbin, R., 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22, 549–556. https://doi.org/10.1101/gr.126953.111

Singtonat, S., Osathanunkul, M., 2015. Fast and reliable detection of toxic Crotalaria spectabilis Roth. in Thunbergia laurifolia Lindl. herbal products using DNA barcoding coupled with HRM analysis. BMC Complementary and Alternative Medicine 15, 162. https://doi.org/10.1186/s12906-015-0692-6

Soininen, E.M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., Brysting, A.K., Sønstebø, J.H., Ims, R.A., Yoccoz, N.G., Taberlet, P., 2009. Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Frontiers in Zoology 6, 16. https://doi.org/10.1186/1742-9994-6-16

Spencer, P.S., Berman, F., 2003. Plant toxins and human health., in: D'Mello, J.P.F. (Ed.), Food Safety: Contaminants and Toxins. CABI, Wallingford, pp. 1–23. https://doi.org/10.1079/9780851996073.0001

Speranskaya, A.S., Krinitsina, A.A., Shipulin, G.A., Khafizov, K.F., Logacheva, M.D., 2018. High-Throughput Sequencing for the Authentication of Food Products: Problems and Perspectives. Russ J Genet 54, 1003–1012. https://doi.org/10.1134/S1022795418090132

Srirama, R., Senthilkumar, U., Sreejayan, N., Ravikanth, G., Gurumurthy, B.R., Shivanna, M.B., Sanjappa, M., Ganeshaiah, K.N., Uma Shaanker, R., 2010. Assessing species admixtures in raw drug trade of Phyllanthus, a hepato-protective plant using molecular tools. Journal of Ethnopharmacology 130, 208–215. https://doi.org/10.1016/j.jep.2010.04.042

Staats, M., Arulandhu, A.J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T.W., Kok, E., 2016. Advances in DNA metabarcoding for food and wildlife forensic species identification. Anal Bioanal Chem 408, 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

Stoeckle, M.Y., Gamble, C.C., Kirpekar, R., Young, G., Ahmed, S., Little, D.P., 2011. Commercial Teas Highlight Plant DNA Barcode Identification Successes and Obstacles. Sci Rep 1, 42. https://doi.org/10.1038/srep00042

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., Willerslev, E., 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35, e14. https://doi.org/10.1093/nar/gkl938

Taboada, L., Sánchez, A., Sotelo, C.G., 2017. A new real-time PCR method for rapid and specific detection of ling (Molva molva). Food Chemistry 228, 469–475. https://doi.org/10.1016/j.foodchem.2017.01.117

Taylor, H.R., Harris, W.E., 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12, 377–388. https://doi.org/10.1111/j.1755-0998.2012.03119.x

Techen, N., Parveen, I., Pan, Z., Khan, I.A., 2014. DNA barcoding of medicinal plant material for identification. Current Opinion in Biotechnology, Analytical biotechnology 25, 103–110. https://doi.org/10.1016/j.copbio.2013.09.010

The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815. https://doi.org/10.1038/35048692

Tillmar, A.O., Dell'Amico, B., Welander, J., Holmlund, G., 2013. A Universal Method for Species Identification of Mammals Utilizing Next Generation Sequencing for the Analysis of DNA Mixtures. PLoS One 8, e83761. https://doi.org/10.1371/journal.pone.0083761

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13, 36–46. https://doi.org/10.1038/nrg3117

Valentini, A., Pompanon, F., Taberlet, P., 2009. DNA barcoding for ecologists. Trends in Ecology & Evolution 24, 110–117. https://doi.org/10.1016/j.tree.2008.09.011

Villa, C., Costa, J., Gondar, C., Oliveira, M.B.P.P., Mafra, I., 2018. Effect of food matrix and thermal processing on the performance of a normalised quantitative real-time PCR approach for lupine (Lupinus albus) detection as a potential allergenic food. Food Chemistry 262, 251–259. https://doi.org/10.1016/j.foodchem.2018.04.079

Waiblinger, H.-U., Boernsen, B., Näumann, G., Koeppel, R., 2014. Ring trial validation of single and multiplex real-time PCR methods for the detection and quantification of the allergenic food ingredients sesame, almond, lupine and Brazil nut. J. Verbr. Lebensm. 9, 297–310. https://doi.org/10.1007/s00003-014-0868-x

Wallace, L.J., Boilard, S.M.A.L., Eagle, S.H.C., Spall, J.L., Shokralla, S., Hajibabaei, M., 2012. DNA barcodes for everyday life: Routine authentication of Natural Health Products. Food Research International 49, 446–452. https://doi.org/10.1016/j.foodres.2012.07.048

Wiel, C.C.M.V.D., Schoot, J.V.D., Valkenburg, J.L.C.H.V., Duistermaat, H., Smulders, M.J.M., 2009. DNA barcoding discriminates the noxious invasive plant species, floating pennywort (Hydrocotyle ranunculoides L.f.), from non-invasive relatives. Molecular Ecology Resources 9, 1086–1091. https://doi.org/10.1111/j.1755-0998.2009.02547.x

Wilcox, T.M., Carim, K.J., McKelvey, K.S., Young, M.K., Schwartz, M.K., 2015. The Dual Challenges of Generality and Specificity When Developing Environmental DNA Markers for Species and Subspecies of Oncorhynchus. PLOS ONE 10, e0142008. https://doi.org/10.1371/journal.pone.0142008

Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84, 9054–9058. https://doi.org/10.1073/pnas.84.24.9054

Wong, E.H.-K., Hanner, R.H., 2008. DNA barcoding detects market substitution in North American seafood. Food Research International 41, 828–837. https://doi.org/10.1016/j.foodres.2008.07.005

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Yao, H., Song, J.-Y., Ma, X.-Y., Liu, C., Li, Y., Xu, H.-X., Han, J.-P., Duan, L.-S., Chen, S.-L., 2009. Identification of Dendrobium Species by a Candidate DNA Barcode Sequence: The Chloroplast psbA-trnH Intergenic Region. Planta Med 75, 667–669. https://doi.org/10.1055/s-0029-1185385

Ye, S.H., Siddle, K.J., Park, D.J., Sabeti, P.C., 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178, 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Yu, M., Jiao, L., Guo, J., Wiedenhoeft, A.C., He, T., Jiang, X., Yin, Y., 2017. DNA barcoding of vouchered xylarium wood specimens of nine endangered Dalbergia species. Planta 246, 1165–1176. https://doi.org/10.1007/s00425-017-2758-9

Zhang, J., Wider, B., Shang, H., Li, X., Ernst, E., 2012. Quality of herbal medicines: Challenges and solutions. Complementary Therapies in Medicine 20, 100–106. https://doi.org/10.1016/j.ctim.2011.09.004

Zhang, L., Lu, H.H.S., Chung, W., Yang, J., Li, W.-H., 2005. Patterns of segmental duplication in the human genome. Mol Biol Evol 22, 135–141. https://doi.org/10.1093/molbev/msh262

Zhang, N., Erickson, D.L., Ramachandran, P., Ottesen, A.R., Timme, R.E., Funk, V.A., Luo, Y., Handy, S.M., 2017. An analysis of Echinacea chloroplast genomes: Implications for future botanical identification. Sci Rep 7, 216. https://doi.org/10.1038/s41598-017-00321-6

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J., Huang, Q., 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. Gigascience 2, 4. https://doi.org/10.1186/2047-217X-2-4

ACKNOWLEDGMENTS

First of all, I am very grateful to my supervisor Maido Remm, who welcomed me to the bioinformatics workgroup. He has always supported and motivated me to move forward and learn new things, always being available for guidance and giving me enough freedom to make my own choices. I am proud to be a member of this bioinformatics workgroup.

I also thank my colleagues from Institute of Molecular and Cell Biology, especially in the department of bioinformatics, for their support, interesting discussions, good life advice, and funny Christmas parties. I've never been bored with you, friends.

I would also like to thank my colleagues from Competence Centre on Health Technologies for inspiring conversations and lunch breaks. I like your positive and "can do" attitude.

I thank my children for their understanding and patience during the periods when I have been busy with seemingly endless work: studying, reading, preparing for the presentations, writing articles, applications, reports, dissertation etc.

And special thanks to my friends and family who have never understood what kind of strange work I do or whether I will stop studying one day, but still always supported me.

Thank you all!

PUBLICATIONS

CURRICULUM VITAE

Name:	Kairi Raime
Date of birth:	08.06.1982
E-mail:	kairi.raime@ut.ee

Education

2015–... PhD studies, Gene technology (bioinformatics), Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia

2012–2015 Master´s studies (MSc), Gene technology (bioinformatics), Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia

2004–2007 Master´s studies (MSc, *magister scientiarum*), Botany and Ecology (plant ecology and ecophysiology), Institute of Botany and Ecology (old name), Faculty of Science and Technology, University of Tartu, Estonia

2000–2004 Bachelor's studies, Biology, Institute of Botany and Ecology (old name), Faculty of Science and Technology, University of Tartu, Estonia

Professional employment:

08.06.2020–... Competence Centre on Health Technologies, Research Fellow of Bioinformatics

01.11.2012–... University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology, Research fellow of Bioinformatics

2006–2012 Asper Biotech, Researcher

2003–2006 Tartu Emajõe School, Teacher of mathematics

Publications:

Raime, Kairi; Krjutškov, Kaarel; Remm, Maido (2020). Method for the Identification of Plant DNA in Food Using Alignment-Free Analysis of Sequencing Reads: A Case Study on Lupin. Frontiers in Plant Science, 11, 646. https://doi.org/10.3389/fpls.2020.00646.

Raime, Kairi; Remm, Maido (2018). Method for the Identification of Taxon-Specific k-mers from Chloroplast Genome: A Case Study on Tomato Plant (*Solanum lycopersicum*). Frontiers in Plant Science, 9, 6. https://doi.org/6.10.3389/fpls.2018.00006.

Kõressaar, Triinu; Lepamets, Maarja; Kaplinski, Lauris; Raime, Kairi; Andreson, Reidar; Remm, Maido (2018). Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics, 34(11), 1937−1938. https://doi.org/10.1093/bioinformatics/bty036.

Laidre, Piret; Soplepmann, Jaan; Uibo, Oivi; Raime, Kairi; Yakoreva, Maria; Mirka, Gerli; Roomere, Hanno; Õunap, Katrin. (2015). Perekondlik adenomatoosne polüpoos: ülevaade ja ühe perekonna haigusjuht. Eesti Arst, 94 (1), 38−43.

Walia, S.; Fishman, GA.; Zernant-Rajang, J.; Raime, K.; Allikmets, R. (2008). Phenotypic expression of a PRPF8 gene mutation in a Large African American family. Archives of Ophthalmology, 126 (8), 1127−1132. https://doi.org/10.1001/archopht.126.8.1127.

Supervised dissertations:

Johanna-Stina Idla, bachelor's degree, 2021, (sup) Kairi Raime, Species-specificity analysis of PCR primers designed to detect nosematosis-causing microsporidia Nosema apis and Nosema ceranae in bees, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology.

Maria Repson, bachelor's degree, 2021, (sup) Kaarel Krjutškov and Kairi Raime, Liquid biopsy-based SNP-genotyping in precision and personal medicine, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology

Siimo Kangruoja, Master's Degree, 2020, (sup) Kairi Raime, Detection of modified organisms from raw sequencing data using *k*-mers of specific length, University of Tartu, Faculty of Science and Technology, Institute of Ecology and Earth Sciences

ELULOOKIRJELDUS

Nimi:	Kairi Raime
Sünniaeg:	08.06.1982
E-post:	kairi.raime@ut.ee

Haridustee

2015–... PhD õpingud, Geenitehnoloogia eriala (bioinformaatika suund), Molekulaar- ja rakubioloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti

2012–2015 Magistrikraad (MSc), Geenitehnoloogia eriala (bioinformaatika suund), Molekulaar- ja rakubioloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti

2004–2007 Magistrikraad (MSc, *magister scientiarum*), Botaanika ja ökoloogia eriala (taimeökoloogia ja ökofüsioloogia suund), Botaanika ja ökoloogia instituut (endine nimi), Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti

2000–2004 Bakalaureusekraad, Bioloogia eriala, Botaanika- ja ökoloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti

Teenistuskäik:

08.06.2020–... Tervisetehnoloogiate Arenduskeskus AS, bioinformaatika teadur

01.11.2012–... Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut, bioinformaatika õppetool, bioinformaatika teadur

2006–2012 Asper Biotech, teadur

2003–2006 Tartu Emajõe Kool, matemaatikaõpetaja

Teaduspublikatsioonid:

Raime, Kairi; Krjutškov, Kaarel; Remm, Maido (2020). Method for the Identification of Plant DNA in Food Using Alignment-Free Analysis of Sequencing Reads: A Case Study on Lupin. Frontiers in Plant Science, 11, 646−646. https://doi.org/10.3389/fpls.2020.00646.

Raime, Kairi; Remm, Maido (2018). Method for the Identification of Taxon-Specific k-mers from Chloroplast Genome: A Case Study on Tomato Plant (Solanum lycopersicum). Frontiers in Plant Science, 9, 6. https://doi.org/10.3389/fpls.2018.00006.

Kõressaar, Triinu; Lepamets, Maarja; Kaplinski, Lauris; **Raime, Kairi**; Andreson, Reidar; Remm, Maido (2018). Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics, 34 (11), 1937−1938. https://doi.org/10.1093/bioinformatics/bty036.

Laidre, Piret; Soplepmann, Jaan; Uibo, Oivi; **Raime, Kairi**; Yakoreva, Maria; Mirka, Gerli; Roomere, Hanno; Õunap, Katrin. (2015). Perekondlik adenomatoosne polüpoos: ülevaade ja ühe perekonna haigusjuht. Eesti Arst, 94 (1), 38−43.

Walia, S.; Fishman, GA.; Zernant-Rajang, J.; **Raime, K.**; Allikmets, R. (2008). Phenotypic expression of a PRPF8 gene mutation in a Large African American family. Archives of Ophthalmology, 126 (8), 1127−1132. https://doi.org/10.1001/archopht.126.8.1127.

Juhendatud väitekirjad:

Johanna-Stina Idla, bakalaureusekraad, 2021, (juh) Kairi Raime, Mesilastel nosematoosi põhjustavate mikrosporiidide Nosema apis ja Nosema ceranae tuvastamiseks disainitud PCR praimerite liigispetsiifilisuse analüüs, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Maria Repson, bakalaureusekraad, 2021, (juh) Kaarel Krjutškov and Kairi Raime, Vedelbiopsiapõhine SNP-genotüpiseerimine täppis- ja personaalmeditsiinis, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Molekulaar- ja rakubioloogia instituut.

Siimo Kangruoja, magistrikraad, 2020, (juh) Kairi Raime, Geneetiliselt muundatud taimede tuvastamine sekveneerimise toorlugemitest kasutades kindla pikkusega *k*–meere, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Ökoloogia ja maateaduste instituut.

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

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

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

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

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

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

106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **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.