

KESSY ABARENKOV

PlutoF – cloud database and
computing services supporting
biological research



TARTU UNIVERSITY PRESS

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

Dissertation was accepted for the commencement of the degree of *Doctor philosophiae* in Botany and Mycology at the University of Tartu on August 26 by the Scientific Council of the Institute of Ecology and Earth Sciences University of Tartu.

Supervisor: Prof. Urmas Kõljalg, University of Tartu, Estonia

Opponent: Prof. Peter Dawyndt, Ghent University, Belgium

Commencement: Room 1019, 14A Ravila Street, Tartu, on 9 November 2011 at 9.15 a.m.

Publication of this thesis is granted by the Institute of Ecology and Earth Sciences, University of Tartu and by the Doctoral School of Earth Sciences and Ecology created under the auspices of European Social Fund.



European Union
European Social Fund



Investing in your future

ISSN 1024–6479

ISBN 978–9949–19–854–2 (trükis)

ISBN 978–9949–19–855–9 (PDF)

Autoriõigus Kessy Abarenkov, 2011

Tartu Ülikooli Kirjastus

www.tyk.ee

Tellimus nr 616

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS	6
LIST OF ABBREVIATIONS	8
LIST OF TERMS AND DEFINITIONS	9
INTRODUCTION	10
THE AIMS OF MY THESIS	14
MATERIALS AND METHODS	15
PlutoF cloud	15
UNITE	15
Biodiversity data	16
RESULTS	18
PlutoF cloud	18
UNITE	21
Biodiversity data	22
Usage statistics	26
DISCUSSION	29
CONCLUSIONS	32
REFERENCES	33
SUMMARY IN ESTONIAN	36
ACKNOWLEDGEMENTS	39
PUBLICATIONS	43

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications referred to in the text by their Roman numerals:

- I. Kõljalg U, Larsson K-H, **Abarenkov K**, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vrålstad T, Ursing BM. 2005. UNITE: A database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* 166 (3): 1063–1068.
- II. Nilsson RH, Ryberg M, Kristiansson E, **Abarenkov K**, Larsson K-H, Kõljalg U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: A fungal perspective. *PLoS ONE* 1: e59.
- III. **Abarenkov K**, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U. 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* 186 (2): 281–285.
- IV. Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, **Abarenkov K**. 2010. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology* 3 (4): 284–287.
- V. Nilsson RH, **Abarenkov K**, Veldre V, Nylinder S, De Wit P, Brosche S, Alfredsson JF, Ryberg M, Kristiansson E. 2010. An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources* 10 (6): 1076–1081.
- VI. **Abarenkov K**, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Proulx M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U. 2010. Plutof—a web based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. *Evolutionary Bioinformatics* 6: 189–196.

Published papers are reproduced with the permission of the publishers.

Author's contribution to each article

	I	II	III	IV	V	VI
Idea and design	+	-	+	+	+	+
Software development	+	-	+	-	-	+
Software testing/data analysis	+	+	+	+	+	+
Writing	+	-	+	-	+	+

LIST OF ABBREVIATIONS

ABCD	Access to Biological Collections Data
BioCASE	the Biological Access Service for Europe
ENA	European Nucleotide Archive
GBIF	Global Biodiversity Information Facility
GSC	Genomic Standards Consortium
INSDC	International Nucleotide Sequence Database Collaboration
ITS	Internal Transcribed Spacer
LSU	large subunit
MCL	Microbiological Common Language
MVC	Model View Controller
rDNA	ribosomal deoxyribonucleic acid
TDWG	Taxonomic Databases Working Group

LIST OF TERMS AND DEFINITIONS

- Biodiversity data – covering all types of data utilized in ecology, genetics and taxonomy.
- Biological sample – any DNA, tissue, specimen, soil, water, air, etc. sample which includes biological material.
- Cloud database – database hosting data that consists of virtually and/or physically separated data units but can be browsed, searched and analyzed all together.
- Cloud computing – model for providing computing resources (e.g. computer power and storage, applications, services) from a server that are executed and managed by a client's web browser.
- PlutoF cloud – system consisting of PlutoF cloud database for storing biodiversity data, web-based workbench for managing and analyzing the data, and several public web pages for accessing subsets of the data.
- PlutoF workbench – web interface for managing and analyzing biodiversity data.
- Taxon occurrence – occurrence of a living organism in nature documented by one-time observation, specimen in collection or any other biological sample.

INTRODUCTION

The term “biodiversity informatics”, probably first mentioned by the Canadian Biodiversity Informatics Consortium in 1992 (<http://www.bgbm.org/BioDivInf/TheTerm.htm>), was introduced to describe informatics tools created for collecting, storing, displaying and analyzing biodiversity data. It grew out from and is strongly overlapping with the fields of molecular bioinformatics and environmental informatics, differing from the latter two by the nature of the data it deals with. Whilst bioinformatics is more concentrated on collecting and analyzing molecular data (genomic, proteomic), and environmental informatics combining database systems, geographic information systems, and simulation modeling to create applications for environmental research and protection, biodiversity informatics has found its niche in linking molecular data of an organism to other biodiversity metadata such as taxon name, its placement in classification system, locality, habitat description, interactions with other organisms (host, substrate), category of threat, etc.

The actual development of biodiversity informatics began after the publication of the OECD Report of the Working Group of Biological Informatics in 1999. This report focused on the problems of developing biodiversity informatics and proposed a plan for creating the Global Biodiversity Information Facility (GBIF) which has become the largest global web portal today for storing and distributing primary biodiversity data. GBIF develops software tools for collecting data from joined institutions and provides search engines for querying and visualizing the data. It supports the development and maintenance of regional/national databases by gathering primary biodiversity data from source databases and linking back to where the original data lies. As of August 2011, 56 countries have joined the GBIF network, including Estonia.

Both the advances in molecular biology and information technology in the last two decades have shaped the landscape of biodiversity informatics in a direction of creating a large number of database-related software tools (both standalone and web-based) for managing these huge amounts of biodiversity data. The development of small scattered “inhouse” databases was soon realized not to correspond to the users community expectations in 1) being easily found and used by potential users; and 2) sharing the same data standards (so that different datasets were comparable). To overcome these problems, the development of global, interoperable systems was initiated starting with GBIF (1999) and followed by BioCASE (2001–2004), Barcode of Life Project (2004), Encyclopedia of Life (2008), and others. GBIF and BioCASE have somewhat overlapping objectives. However, GBIF has a global scope, focusing on digitized collections, while BioCASE is interested in European collections, not necessarily supported by databases. GBIF currently focuses on mobilizing primary species occurrence data (specimen and observation records) and creating an Electronic Catalogue of Named Organisms. Barcode of Life Project is an international collaboration to build a DNA barcode reference library. With their developed informatics workbench The Barcode of Life Data System, BOLD

(Ratnasingham & Hebert, 2007), their goal for the next 5 years is to barcode 5 million specimens representing 500 000 species. Encyclopedia of Life is an online collaborative encyclopedia compiled from existing databases and with an ambition to document all 1.9 million living species known to science.

Biodiversity data can be used by a wide range of scientific areas such as answering ecological and biogeographical questions (species distributions, interactions between and co-evolution of species), identifying threatened species and deciding upon actions for their protection, measuring of environmental impacts, etc. The field of biodiversity informatics currently focuses on the following aspects:

1. developing tools for data quality control. With massive amounts of primary biodiversity data already present and the continuous addition of new data into large databases such as GBIF and INSDC (Benson et al., 2006), the issue of data quality has come up in recent years (Bridge et al., 2003; Guralnick et al., 2007);
2. developing biodiversity standards to standardize data fields in different systems for networking all databases holding biodiversity data. The most important organization developing and promoting standards for exchange of biological/biodiversity data is the Biodiversity Information Standards (TDWG). The two primary standards that TDWG develops are DarwinCore and ABCD. While the main focus of both standards was initially the terminology associated with biological collection data, recently they have been supplemented with extensions for other fields such as geosciences and DNA data. Other important collaborations involved in developing standards for molecular biology include the Genomic Standards Consortium (GSC) that create standards for genomics and genome descriptions (Yilmaz et al., 2011), and the microbiology workgroup developing Microbiological Common Language (MCL) for standardizing the electronic exchange of meta-information about microorganisms (Verslyppe et al., 2010);
3. developing sophisticated search engines, data visualization and analysis tools (e.g., the Barcode Of Life Data System (BOLD), Interactive Tree Of Life (iTOL) project, geographic information systems) in addition to the relatively simple software for capturing, storing and displaying the data that was the primary purpose of biodiversity informatics tools in initial large collaborative projects (GBIF).

Identification of fungal species in ecological studies is nowadays most commonly DNA-based, because 1) fungal fruit-bodies for morphological identification are not that frequently formed or these are missing in some fungal taxa; 2) for some groups of species, morphological characteristics fall short for distinguishing between closely related taxa; 3) DNA-based identification is in many cases faster and cheaper, and does not necessarily require expertise in specific fungal groups. Until recently, the traditional Sanger sequencing was the most commonly used DNA sequencing method for identifying fungal species in

biological (soil, air, wood, gut) samples. During the last few years however the massively parallel 454 pyrosequencing method (Roche) has been used successfully for assessing the fungal communities *in situ* (Bueé et al., 2009; Jumpponen and Jones, 2009; Öpik et al., 2009; Tedersoo et al., 2010). 454 sequencing enables to process a much higher number of samples to a greater depth at the same time and cost. However, this technology has its inherent shortcomings such as 1) the length of DNA fragments that it enables to sequence, is fairly short (400–500 bp as of August, 2011); and 2) the quality of DNA sequences it produces is affected by a large number of base reading errors (Huse et al., 2007). These problems will certainly be solved in the future and the improved pyrosequencing methods will definitely gain wider use in the community.

Although not officially approved as a DNA barcode, the ITS region (ITS1-5.8S-ITS2) of the fungal ribosomal DNA is the most widely used marker for DNA-based identification of fungi. It has been used in ecological studies already for more than 15 years. The ITS region became popular among fungal ecologists, because 1) each copy of DNA includes up to a few hundred ITS copies which makes it easy to amplify from very small samples; and 2) it is variable enough to differentiate between closely related species in most fungal taxa. The length of the ITS regions usually varies between 450–650 bp, whereas the sequence length gained in pyrosequencing analysis is now up to 700–1000 bp. Full-length ITS sequences will improve the accuracy of identification of unknown fungal taxa.

Using DNA-based methods for the identification of organisms requires availability of an inclusive reference database for sequence comparison. Ideally, the reference database should meet the requirements of 1) featuring satisfactory taxonomic coverage of sequences; and 2) including only sufficiently annotated sequences of good quality that originate from vouchered specimens identified by an expert. The most widely used nucleotide sequence reference database for all organisms has been the consortium International Nucleotide Sequence Database Collaboration (INSDC) which is formed by DDBJ (Japan), NCBI (USA) and ENA (Europe) databases. Most scientific journals require that the DNA sequences published in an article be submitted to INSDC database. In addition to sequence data INSDC also offers informatics tools for species identification through various BLAST algorithm-based similarity searches (Altschul et al. 1997) and for constructing phylogenetic trees.

Although INSDC is the most inclusive and widely used database today, it does not correspond to either of the two above requirements: 1) less than 1% of the estimated 1.5 million species of fungi is sequenced for the ITS region, the most widely used locus for species identification of fungi (Nilsson et al., 2005; Mueller and Smith, 2007); and 2) a large proportion of fungal ITS sequences deposited in INSDC are either of low quality, misidentified or poorly annotated. Data fields in INSDC are often unstandardized with important metadata in unstructured formats or missing at all, and missing or wrong annotations are seldom complemented by the original submitters (Nilsson et al., 2006). Third-

party annotations to improve the quality of INSDC data and possibility to alert its users of the misidentifications are suggested by the research community (Bardotondo et al., 2008), but these options are not yet implemented in the INSDC.

The taxonomic coverage of fungi in INSDC is expected to rise with the new technology, pyrosequencing, for mapping entire fungal communities in diverse biological samples, and with sequencing fungal fruitbodies already present in herbaria but for which no DNA sequence is currently available (Brock et al., 2009; Nagy et al., 2011; Rosling et al., 2011). But for reliable DNA-based identification of organisms the dataset currently in INSDC needs to be revisited and annotated so that the new sequences identified and deposited in the future would not carry on the mistakes it contains. There have been several efforts to achieve this by developing more accurate but less inclusive, curated databases such as SILVA (Pruesse et al., 2007), Greengenes (DeSantis et al., 2006), MaarjAM (Öpik et al., 2010) and UNITE (Kõljalg et al., 2005; Abarenkov et al., 2010).

The analysis of such a huge amount of sequence data and their associated metadata requires biodiversity informatics tools to be developed as has been done by Ludwig et al. 2004 (ARB), Schloss et al. 2009 (mothur), and Caporaso et al. 2010 (QIIME). The critical requirement in all of them is the presence of a good reference dataset consisting of sequences that are correctly identified.

THE AIMS OF MY THESIS

The overall aim of the thesis was to develop technologies for managing, editing and analyzing biodiversity data. I focused on the following topics:

- 1) developing rDNA ITS based identification tools for fungi;
- 2) developing public web-based system for molecular identification of fungi;
- 3) developing system for third-party annotations of all publicly available fungal ITS sequences;
- 4) developing e-infrastructure for biodiversity database services;
- 5) developing web-based workbench for management of biodiversity data.

Uploading data into PlutoF cloud database (except the fungal rDNA sequences) was not the primary aim of this study. However, the number of PlutoF cloud users/databases and amount of uploaded data shows the applicability of the developed solutions and is therefore covered in the thesis as well.

MATERIALS AND METHODS

PlutoF cloud

PlutoF workbench runs on two quad-core 64-bit Linux servers erast.ut.ee and hermes.ut.ee (CentOS 5.2, Apache web server v. 2.2.3) where hermes.ut.ee is used for database replication and sharing data with BioCASE portal. Database management system involves MySQL 5.0.77. PlutoF web interface was built using the following web technologies: PHP (current version 5.3.3), HTML, CSS, AJAX and JavaScript. Software packages of the analysis module were written in Perl (current version 5.8.8). CPU-intensive computations in the analysis module are sent to and carried out at the High Performance Computing Center of the University of Tartu (<http://www.hpc.ut.ee>). PlutoF workbench is available online at <http://plutof.ut.ee>.

UNITE

UNITE is a database of fungal rDNA ITS sequences comprised of sequence data in PlutoF cloud database and public homepage at <http://unite.ut.ee> for carrying out searches and molecular identification. Sequence data in UNITE can be divided into 3 separate datasets depending on their reference status and origin: 1) UNITE *reference* dataset – high-quality reference ITS sequences isolated from fruit-bodies which are presented with rich metadata and identified by the experts. Data has been added since January 2002. 2) UNITE *Envir.* dataset – high-quality ITS sequences originating from non-specimen biological samples which are presented with rich metadata and submitted by the UNITE work-group members. Data has been added since January 2008. 3) UNITE *INSDC* dataset – database of all fungal rDNA ITS and LSU sequences downloaded from INSDC on a bimonthly basis. To be downloaded from INSDC, the sequences must fill certain quality criteria (Table 1). Data are being added since November 2009.

Table 1. Describing the method of filtering out fungal rDNA ITS and LSU sequences based on their sequence length, classification in “organism” field of GenBank record, and keywords in “title” field of GenBank record.

Region	Length between	Organism	Title	Sequence retrieval string
ITS	140 – 3 000	Fungi, NOT Uncultured Neocallimastigales *	ITS1, ITS2, 5.8S, internal transcribed spacer, internal transcribed spacers, ITS 1, ITS 2	((("Fungi"[Organism] AND (140[SLEN] : 3000[SLEN])) AND (((ITS1[titl] OR ITS2[titl]) OR 5.8S[titl]) OR "internal transcribed spacer"[titl] OR "internal transcribed spacers"[titl] OR "ITS 1" [titl] OR "ITS 2"[titl])) NOT "Uncultured Neocallimastigales"[Organism])
LSU	300 – 10 000	Fungi	LSU, large subunit ribosomal	((("Fungi"[Organism] AND (300[SLEN] : 10000[SLEN])) AND ("LSU"[titl] OR "large subunit ribosomal"[titl]))

* there were > 260 000 sequences flagged as “Uncultured Neocallimastigales” retrieved by search string that we chose to filter out because of their high level of redundancy and origin from a single 454 pyrosequencing study (Liggenstoffer et al., 2010).

Biodiversity data

Biodiversity data in the PlutoF cloud database originates from a number of projects, workgroups and individual researchers reflecting various different fields in biodiversity research starting from managing data in natural history collections and ending with molecular identification of species. The main data holders (workgroups and individual researchers) can be divided into four groups based on their objectives:

1. Natural History Collections – managing herbarium and collection specimens. Main collaborating institutions comprise the University of Tartu Natural History Museum (Museum of Zoology, Museum of Botany), Estonian University of Life Sciences (Fungal herbarium, Fungal culture collection, Department of Plant Protection, Department of Zoology at the Institute of Agricultural and Environmental Sciences) and Tallinn Botanical Garden. Data have been added since April 2007.
2. Estonian Species Registry – keeping the list of species marked as being present in Estonia based on specimen in collection, observation or published literature reference. The project to create Estonian Species Registry was carried out in 2008–2010 in collaboration of the University of Tartu, Estonian University of Life Sciences, Estonian Environment Information Centre and Estonian Naturalists’ Society.
3. Observations – workgroups and individuals adding observations of fungi, plants and animals. This also includes recording taxa in students’ field

courses and creating species lists for protected areas. Data have been added since March 2008.

4. Research groups – several research groups dealing with taxonomy, ecology and biogeography are using PlutoF cloud to manage and analyze their studies, plots and samples that are associated with collection and molecular data. Data have been added since January 2008.

Biodiversity data about Estonia in PlutoF cloud database can be accessed online over the PlutoF cloud database at Estonian eBiodiversity web page (<http://elurikkus.ut.ee>) which has been built using the same web technologies as PlutoF workbench itself.

RESULTS

PlutoF cloud

PlutoF cloud is a “cloud database” with web-based workbench for data management built upon it as a “thin client” (VI). It also includes several public web sites for displaying and searching the data (Figure 1). PlutoF cloud can be thought of as an umbrella for different datasets that can all form individual databases, but can also be linked to each other and treated as one complex system, e.g. classification for keeping the Estonian Species Registry, specimen collections of natural history museums, ecological studies and samples of DNA sequences. In PlutoF cloud these data are stored in a single relational database with a database model consisting of more than 150 tables (Suppl. Item 1 in VI). The database structure is rooted in Taxonomer (Pyle, 2003), but includes substantial modifications to integrate modules for storing multimedia, molecular and ecological data, and analysis results. The current database model supports uploading biodiversity data concerning various taxon occurrences (based on, e.g., specimens, observations or DNA sequences), literature references and scientific collections. PlutoF database structure and workbench features are designed specifically to be successfully used by research groups and individual researchers in the fields of taxonomy and ecology. My group has implemented the hierarchical study/plot/sample model (Figure 1 in VI) that enables users to manage their own projects from sampling design to molecular data analysis. PlutoF database structure is still frequently updated to follow the standards proposed by TDWG and include new modules for, e.g., adding living specimens, laboratory notebooks, etc.

PlutoF workbench features a login system, where user rights and privileges are determined by the username. User rights management is implemented on database level. User can be a member of any number of workgroups, whereas read and write privileges, as well as the actions that the user can perform within this workgroup (for example, manage collection specimens, complement Estonian Species Registry, annotate the INSDC sequence dataset), are determined on the workgroup level as workgroup properties.

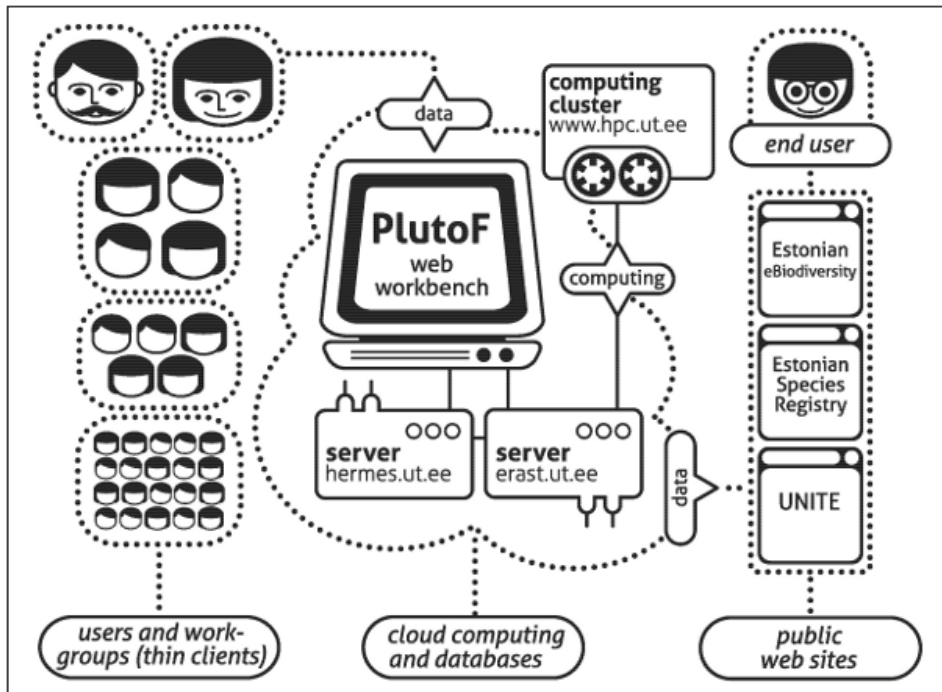


Figure 1. A schematic drawing of the functioning of PlutoF cloud. Compiled by Marie Kõljalg.

On the workbench users have access to their own (either private or public) and workgroup data that has been made available for other workgroup members. These data can be browsed, searched and analyzed together. PlutoF workbench features a clipboard system where data can be sorted out and sent to the clipboard for further processing (eg, download data into files, display plot localities on a map, carry out analysis with molecular data). The analysis module in PlutoF workbench currently comprises software tools for:

- extracting ITS1 and ITS2 subregions of the ITS region from the flanking rDNA genes using the ITS Extractor (IV; VI)
- identifying potentially chimeric ITS sequences through contrasting the respective taxonomic signal of the ITS1 and ITS2 subregions (V; VI)
- clustering sequences using single-linkage clustering at user-defined similarity threshold values with BLASTClust of the BLAST suite (Altschul et al., 1990; VI)
- comparing large query datasets for similarity against sequences in UNITE and INSDC datasets with serial BLAST engine (VI)
- analyzing pyrosequencing datasets of the ITS region using the pyrosequencing pipeline (Tedersoo et al., 2010; VI)

- identifying relevant insufficiently identified sequences in INSDC dataset using an integrated BLAST-based search tool emergencia (Nilsson et al., 2005; VI)

More time and memory demanding analyses are sent to and carried out at the High Performance Computing Center of the University of Tartu (<http://www.hpc.ut.ee>).

Since 14 October 2010, 1 269 analysis runs have been started on PlutoF workbench by 63 distinct users whereas 12 out of top 15 analysis runners are not users from Estonia nor are they closely related to the mycology workgroup at the University of Tartu. The highest number of runs belongs to seriate BLAST search tool (428) followed by the 454 pyrosequencing pipeline (351), ITS Extractor (311), Chimera Checker (118) and BLASTClust (61). The number of analysis runs between October 2010 and July 2011 is shown in Figure 2.

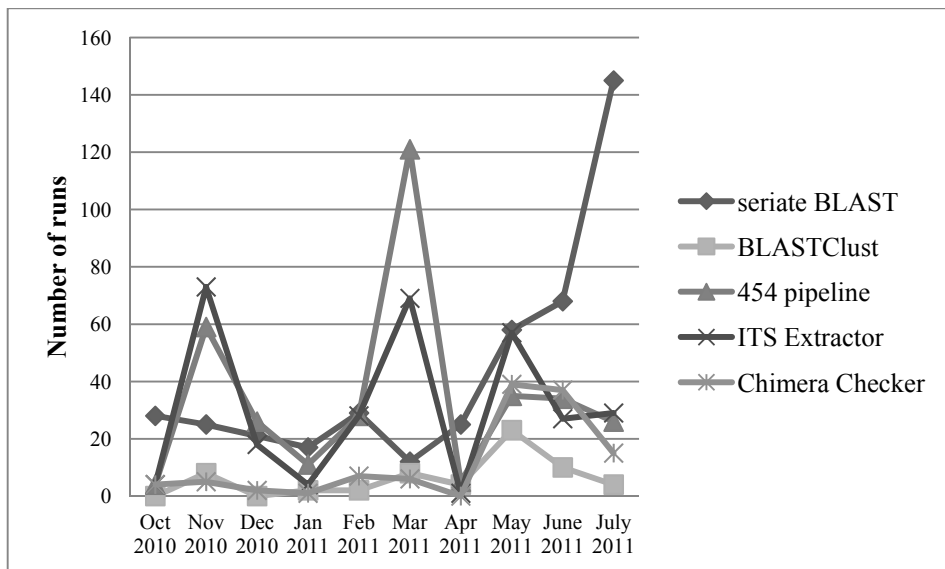


Figure 2. The number of analysis runs started on PlutoF workbench during Oct 2010–July 2011.

As of August 2011, there are 403 PlutoF workbench users belonging to 25 distinct workgroups. Table 2 shows the list of public workgroups which involve more than 10 workgroup members. Complete list of public workgroups and users of PlutoF cloud can be found online at <http://elurikkus.ut.ee/plutof.php?lang=eng&wg=1>.

Table 2. List of public workgroups and their main activities on PlutoF workbench involving more than 10 workgroup members.

Workgroup name	Main activities	Members included
Bird observations	Uploading bird observations	91
Identifications	Using emergencia, UNITE INSDC dataset and analysis module for species identification	79
Estonian fungal collections	Managing specimen data of Estonian mycological collections	59
Sequence annotations	Annotating UNITE INSDC sequences	56
Estonian Species Registry	Managing data for taxa present in Estonia and updating classification	44
UNITE	Adding and editing fungal DNA sequences in the UNITE database	36
Plant root	Managing studies and literature about plant roots	28
Estonian plant collections	Managing specimen data of Estonian plant collections	19
Estonian animal collections	Managing specimen data of Estonian animal collections	19

On average, there are 4–8 workbench users constantly logged in during a working day. The average number of logins per week is 306 (see Figure 5A in Usage statistics section for more details).

UNITE

UNITE is a fungal rDNA ITS sequence database that was originally designed to store high-quality ITS sequences generated from fruiting bodies collected and identified by experts (I, III). The main purpose of the database was to provide the data and tools needed to effectively and reliably identify fungal DNA from environmental samples. The first version of UNITE was made available online in 2003; it enabled users to run simple BLASTn searches and phylogenetic sequence identification using galaxie (Nilsson et al., 2004). When first published in 2005 the UNITE database contained 758 ITS sequences from 455 species and 67 genera of ectomycorrhizal fungi with mainly Baltic-Nordic distribution (Kõljalg et al., 2005).

With the third version of UNITE published in 2010 (III) a number of new features was added:

- UNITE’s initial geographical and ecological restrictions were lifted, and it now also accepts the deposition of unidentified sequences from any biological sample provided that they are of high quality and well annotated;

- New data model was adopted together with PlutoF workbench for uploading and editing the data;
- Serial BLAST engine was developed and made available on public homepage. Serial BLAST for larger datasets and other more computer time and memory demanding tools like 454 pipeline, ITS Extractor and Chimera Checker are available for registered users over the PlutoF workbench;
- Local copy of annotated INSDC dataset was made available through seriate BLAST/blastn tools, and search engines on public UNITE homepage.

As of August 2011, UNITE contains 2 855 reference ITS sequences of 1 116 species from 149 genera. In addition, 159 sequences wait for release to public access (sequences not published in scientific article yet and locked by their submitters). The number of sequences in UNITE *Envir.* dataset (sequences originating from ectomycorrhizal root samples) is currently 1 236 and 2 559 sequences wait for release.

To overcome problems such as misidentifications (we found that up to 20% of fungal rDNA ITS sequences have compromised taxonomic annotations, II), missing and unstandardized metadata and poor quality (II, III) in public gene repositories (INSDC), we decided to use PlutoF cloud to store a regularly updated local copy of fungal rDNA ITS and LSU sequences for which annotations like adding determinations, specifying metadata and quality checks can be made.

As of August 2011, the UNITE *INSDC* dataset consists of 205 798 ITS sequences (either ITS1, 5.8S or ITS2 present) and 29 612 LSU sequences representing 14 104 distinct submissions in INSDC. This dataset is regularly checked for reverse complementary sequences (Nilsson et al., 2010) and chimeric sequences using Chimera Checker (V). Low quality sequences are marked based on either the number of ambiguous nucleotides present in sequence data or by annotators personal opinions. The number of sequences flagged as being chimeric, reverse complementary or of low quality is 680, 1 503 and 2 578 respectively. As of July 2011, UNITE is an ENA LinkOut provider, which means that all the sequences present in INSDC that are also present in UNITE are hyperlinked in ENA.

Biodiversity data

Three main institutions storing their specimen data in PlutoF cloud are the University of Tartu Natural History Museum, Estonian University of Life Sciences and Tallinn Botanical Garden. As of August 2011, the database contained 270 579 specimens from these institutions belonging to six kingdoms of life, with animals and fungi being the best represented (Table 2).

Table 2. The number of databased specimens in Estonian natural history collections grouped by kingdom.

Kingdom	No. of specimens
Plantae	17 021
Animalia	106 226
Fungi	145 025
Chromista	44
Bacteria	1
Protista	2 262
Total	270 579

The yearly addition of specimens in different institutions can be found in Table 3. As yet, the only bacterium in database is determined as a bacterial infection on a fungal species *Cystoderma amianthinum* (Scop.) Fayod collected by Kadri Pöldmaa (herbarium nr: TU112696). Specimens belonging to Chromista are deposited in the Estonian University of Life Sciences fungal collections while specimens belonging to Protista have been deposited in the collections of all three institutions.

Table 3. Yearly addition of databased specimens in Estonian natural history collections since 2007 (January–July for 2011).

Institution	University of Tartu			Estonian University of			Tallinn Botanical Garden		
	Natural History Museum	Plantae	Fungi	Animalia	Plantae	Fungi	Animalia	Plantae	Fungi
2007	0	502	4	1 907	0	3 113	0	0	0
2008	11 265	8 472	40 512	33 289	685	43 493	0	2	7 852
2009	1 118	2 156	2 416	11 290	0	11 170	0	0	0
2010	24 113	2 642	7 901	17 585	1	15 571	0	134	399
2011	2 120	2 094	4 152	3 539	1	7 355	0	332	1 087
Total	38 616	15 866	54 985	67 610	687	80 702	0	468	9 338

Creation of the PlutoF workbench started in 2007, when the first version was made available through a simple web interface with database structure able to store classification and primary specimen data. Only a modest number of specimens were added in 2007. In 2008 data was transferred from various smaller databases (e.g. MS Access) and text files (e.g. MS Excel, CSV) into the new system in an automated way. In the year 2009, a new version of PlutoF workbench was released and the majority of specimens (except certain groups belonging to Animalia that were uploaded from MS Excel template files) are being added through the workbench by its users since then.

Estonian Species Registry is a list of species and other taxa found in Estonia. As of August 2011, the total number of species based on collection specimens, human observations and literature references totals 24 672 (with synonymy taken into account, Table 4). As can be seen from Table 4, the total number of species in Estonian Species Registry is mostly contributed by literature references (94% of species described), whereas specimens in collections and human observations cover only 44% and 7% of the total number of species, respectively. During the years 2008–2010, plant, animal and fungal taxonomists uploaded and verified literature references where, species were marked as being present in Estonia. The overall process of adding taxon reference-based occurrences in PlutoF cloud is presented on Figure 3. Based on literature references almost half of the recorded species in Estonia belong to the kingdom Animalia (11 187) followed by Fungi (6 227), Plantae (4 041), various protists (1 511) and bacteria (258). One could assume that adding the next reference into the database would potentially add 14 new animal species known to Estonia (taken that 407 references currently in database mark the 11 187 unique species and 21 527 taxon occurrences in total).

Table 4. Total number of Estonian species in each kingdom based on literature references, specimens in collections and human observations.

	Total number of species	Based on literature references	Based on specimens in collections	Based on human observations
Animalia	11 621	11 187	4 679	501
Plantae	4 111	4 041	1 191	734
Fungi	7 150	6 227	4 937	488
Protista	1 532	1 511	135	0
Bacteria	258	258	0	0
Total	24 672	23 224	10 942	1 723

As of August 2011, there are 91 277 public species observations in PlutoF cloud database, of which 85 825 belong to animals (85 593 bird observations), 3 029 to plants and 2 423 to fungi. The growth rate of adding observations through PlutoF workbench is shown in Figure 4. The number of plant observations was growing during the years 2008–2009 when PlutoF was used to record species lists in botanical field courses. The number of fungal observations has grown steadily from 2008 to present; it is currently used for recording taxa in mycological field courses and forays, and for creating species lists for protected areas. The number of animal observations began its fast growth in early 2010 largely due to the bird observer community who started to actively use PlutoF cloud for their recordings.

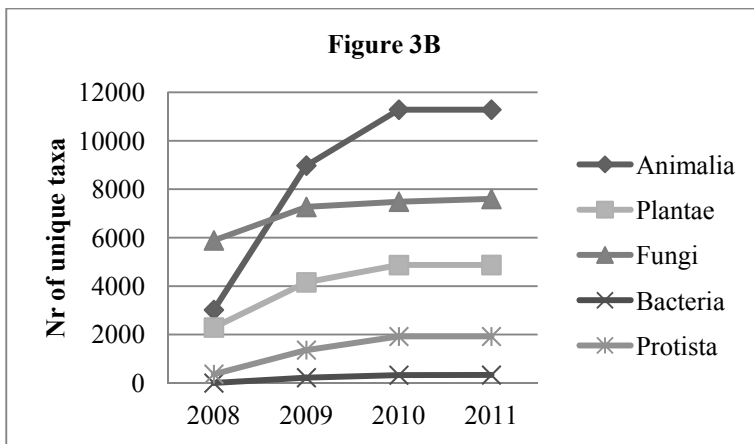
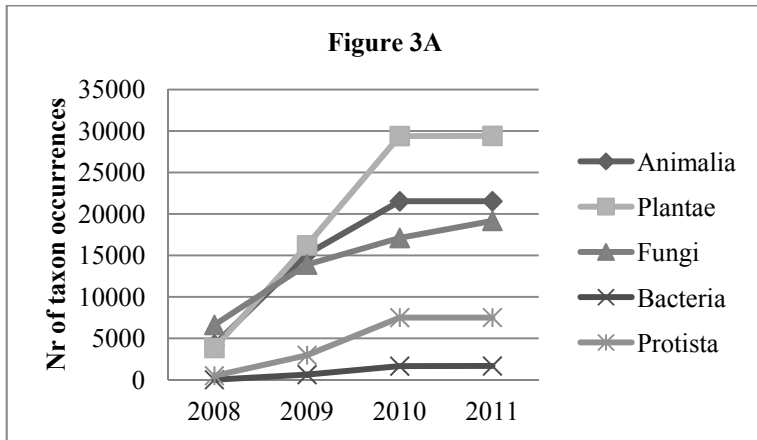


Figure 3. Cumulative yearly addition of A) taxon occurrences; and B) unique taxa in Estonia based on literature references.

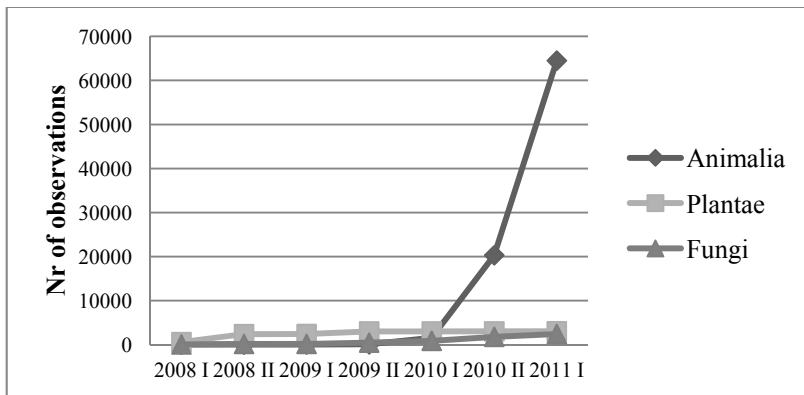


Figure 4. Addition of species observations during the years 2008–2011 in three kingdoms.

Estonian Species Registry can be browsed and searched online at the Estonian eBiodiversity web page (<http://elurikkus.ut.ee/index.php?lang=eng>). For each species known to Estonia, its placement in classification system, synonymes, data from Estonian Red List of Threatened Species, reference data about species occurrence in Estonia, specimens in scientific collections, human observations, public gene sequences, photos and distribution map of all databased records with geo-coordinates are shown (where available, Suppl. Item 1). Specimens of animals and fungi in Estonian natural history collections can be further browsed and searched at the National database of Estonian animal collections (http://unite.ut.ee/eesti_loomakogud/) and National database of Estonian fungal collections (<http://unite.ut.ee/EestiLiigid/>), respectively, where the search is not limited to specimens collected from Estonia but allows to browse all specimens collected from 152 distinct countries.

Usage statistics

To describe latest weekly usage of PlutoF, UNITE, and Estonian eBiodiversity web pages, Piwik 1.5.1 (<http://piwik.org/>) statistics for time period of 12–18 August 2011 was used. For the last 6 months usage of these web sites, PHPCounter 7.2 log files for time period during Feb 17 2011 – July 28 2011 were evaluated.

Piwik usage statistics (Table 5) shows that Estonian eBiodiversity web site has the highest number of unique visitors (1 302) while PlutoF workbench is characterized by the highest number of total actions (page views and downloads, 14 284), the highest average number of actions per visit (56.2), and the highest maximum number of actions per visit (931) committed by a fairly low number of visitors (254). The average visit duration of these 3 systems clearly separates PlutoF workbench, an every-day working tool, from UNITE and Estonian eBiodiversity – public web sites with their main focus on displaying biodiversity data for wider audience.

Table 5. The usage overview for PlutoF workbench, UNITE and Estonian eBiodiversity web sites during 1 week.

	PlutoF	UNITE	eBiodiversity
Number of unique visits	254	160	1 302
Number of actions	14 284	420	5 604
Average number of actions per visit	56.2	2.6	4.3
Average visit duration	37 min	4 min 5 sec	3 min 35 sec
Maximum number of actions per visit	931	48	145

The most visited web pages suggested by Piwik usage statistics were selected for displaying the six-month view count variation for these pages. The number of logins to PlutoF workbench averages 306 per week and it has been quite stable during the 6 months period with small decline in the mid-summer period (Figure 5A). The number of visits to species information pages has been stable during the period, but the number of visits to observations page started to rise in the end of March when the bird observation season began. Visits to species search pages and Estonian Species Registry have been quite stable for the first 4 months showing a decline in the summer period.

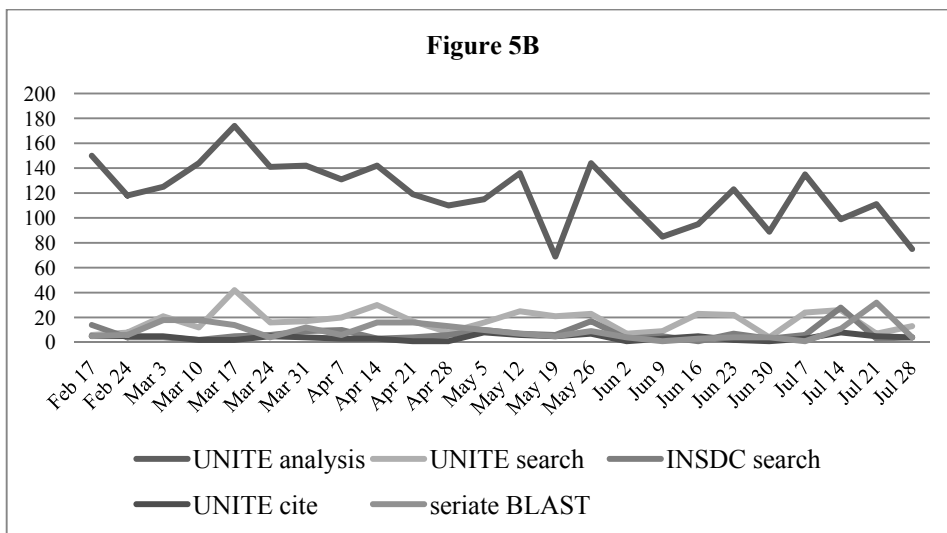
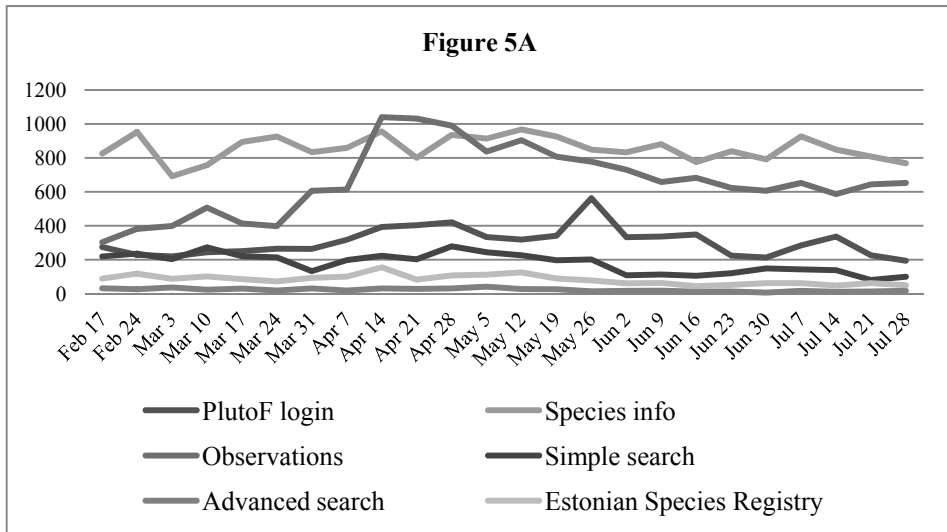


Figure 5. The number of visits to A) PlutoF workbench login and Estonian eBiodiversity pages; and B) UNITE web site's most visited pages with a weekly interval.

On average there were 120 unique hits per week in the UNITE analysis page, followed by UNITE search (17), seriate BLAST (9), INSDC search (7) and citing information (4) pages (Figure 5B). Great difference between the number of visits to analysis and seriate BLAST pages indicates that most users are either not yet aware of the seriate BLAST tool or are using more powerful version available over the PlutoF workbench for registered users.

The proportion of page views by visiting countries for Estonian eBiodiversity and UNITE web pages is given on Figure 6A and Figure 6B. The total number of distinct countries according to Piwik usage statistics during the one week period was 42 for Estonian eBiodiversity, 22 for UNITE homepage, and 12 for PlutoF workbench. Figure 6 reveals that although Estonian eBiodiversity is visited by users from the highest number of distinct countries, the overall proportion of page views from foreign countries is only 23%, while the proportion of page views from visitors outside Estonia for the UNITE homepage amounts 86%, referring to the mostly international usage of the UNITE database.

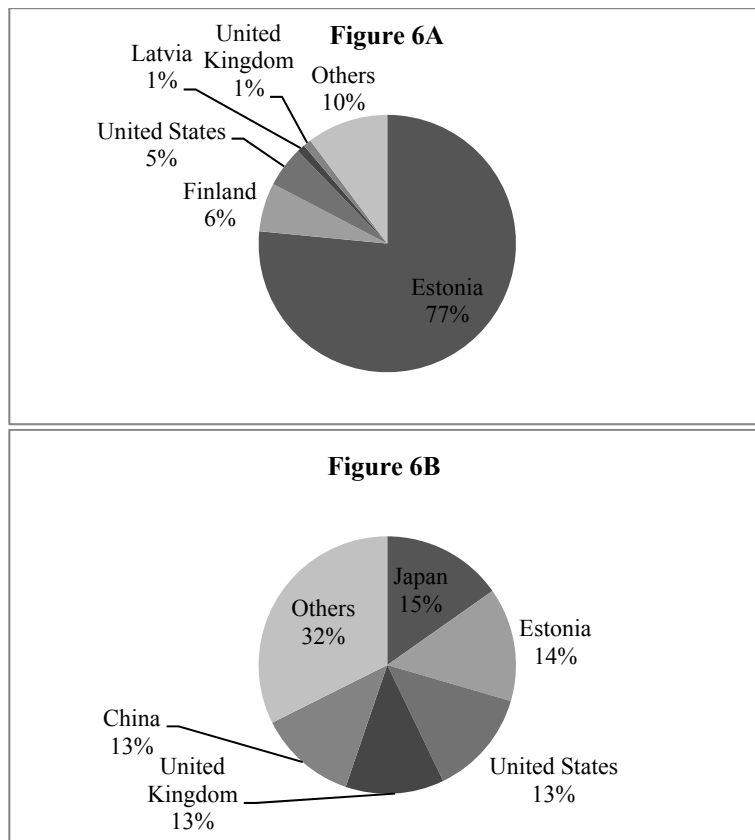


Figure 6. Proportion of A) Estonian eBiodiversity and B) UNITE page views by visiting countries.

DISCUSSION

The overall aim of this thesis was to develop a system for managing and analyzing biodiversity data. Its main focus was on developing: 1) a standardized data model for storing the data relevant to ecological and taxonomical research; 2) a web based workbench for managing the data; 3) analysis tools for molecular identification of fungi; and 4) web pages for public access to these data and analysis tools.

Molecular methods for species identification of fungi became widely used in the end of 1990s (Horton & Bruns, 2001). Initially popular RFLP methods were soon replaced by comparing DNA sequences of certain region for similarity, e.g. DNA sequences from ectomycorrhizal roots were compared for similarity against DNA sequences originating from fungal fruit-bodies with species name present. For DNA-based identification, the availability of a good reference dataset with satisfactory taxonomic coverage is crucial. In the beginning of 2000 this role was filled by the INSDC, whose taxonomic coverage for fungi was still limited and misidentifications were far from rare (Vilgalys, 2003). To enable fast and reliable molecular identification of fungi, we created UNITE (I) – database consisting of rDNA ITS sequences from fungal fruitbodies identified by experts – aiming to fill the gap in INSDC for ectomycorrhizal fungi with Nordic-Baltic distribution. The UNITE database was released on the web in 2003. The first UNITE paper published in 2005 and the high number of citations it has received to date (162, according to SCOPUS citation database as of August 2011) indicates its necessity and broad use by the fungal research community. The UNITE reference dataset has also been used in many “inhouse” analysis tools by several international workgroups, e.g. SCATA for sequence clustering and analysis of tagged amplicons (Durling et al., 2011). As of July 2011, all publicly available UNITE reference sequences as well as the annotated INSDC dataset are available for download as a FASTA file at <http://unite.ut.ee/repository.php>.

Meta-analysis of fungal sequences in INSDC conducted in following years by my group (II; Ryberg et al., 2009) further emphasized the insufficiency of metadata available in INSDC, and the fact that misidentifications were prone to carry on to newly uploaded sequences. Alongside UNITE, which was initially focused on ectomycorrhizal fungi, other group- or gene-specific curated reference databases were developed, e.g. Greengenes for 16S rRNA (DeSantis et al., 2006), SILVA for small and large subunits of rRNA (Pruesse et al., 2007) and MaarjAM for arbuscular mycorrhizal fungi (Õpik et al., 2010).

Ryberg et al. (2008, 2009) and Nilsson et al. (2010A) showed the importance of including insufficiently identified sequences (IIS, sequences without full species-level identification, often named as “unidentified fungus” or “uncultured fungus” originating from non-specimen biological samples) in the taxonomic and ecological analyses. The proportion of IIS in INSDC is growing rapidly – as of November 2010, 42% of ITS sequences in INSDC were insufficiently identified compared to the 27% we reported in 2006 (II). As we

showed then, the IIS and fully identified sequences (FIS, sequences with full species-level identification) form two distinct subsets of the full sequence dataset indicating the presence of possibly high number of yet undescribed taxa among the IIS dataset. These two facts suggest that IIS dataset is an important source of data in studies dealing with biogeography, host-specificity and phylogeny of fungi.

In addition to the insufficient annotations in INSDC, there are also sequence quality issues. As we reported already in 2006, the proportion of sequences with more than 1% IUPAC DNA ambiguities was 1.8% (II). With the number of sequences grown for almost four-fold in the subsequent 5 years, this percentage has remained unchanged. Another problem is the presence of chimeric and reverse complementary sequences that probably accumulate with an increasing number of environmental studies. The proportion of reverse complementary sequences – sequences that are cast backward and in which all purines and pyrimidines are transposed – in INSDC is about 1% as shown by Nilsson et al. (2010B). As we showed in 2010 (V), the estimated proportion of chimeric sequences – sequences that are formed by parts of sequences from 2 or more distinct organisms – is 1.5%. If the reverse complementary sequences are more of an inconvenience for a researcher, then chimeric sequences are a more serious threat in e.g. giving false results in BLAST based similarity search tools and causing higher estimation of species richness when similarity based sequence clustering is used to calculate it.

Due to these shortcomings of INSDC we decided to download and keep a local bimonthly updated copy of fungal rDNA ITS sequences that we would be able to correct and annotate (add determinations, metadata on locality, habitat and interacting taxa, flag chimeric and low quality sequences, etc.) To provide the international working group with the tools to add annotations, we developed a web-based workbench PlutoF (VI) featuring a login system and tools for analyzing molecular data. The comparison of PlutoF workbench with two other software packages for analyzing molecular data, such as mothur and QIIME, showed that the main features distinguishing PlutoF from the latter two are: 1) the possibility to share data within workgroups; 2) the possibility to annotate reference dataset available for the whole research community; 3) advanced search options for the reference dataset; and 4) analysis programs designed specifically for variable fungal ITS sequences.

Sequence data coming from the first massively parallel 454 pyrosequencing studies was exceptional in a way that the sequence length this method allowed to generate was enough for sequencing only a part of ITS1 or ITS2 subregion of the full ITS region. When sequencing either ITS1 or ITS2 and using BLAST algorithm-based similarity searches for identification, the flanking conserved gene regions, depending on their length, will always find matches in sequence databases, even if the ITS1 or ITS2 do not. This makes the identification process more complicated and automatic interpretation of the BLAST results appears problematic. To remove these flanking conserved regions and extract ITS1 and ITS2 of the ITS region, we developed the ITS extractor (IV) which

was later also used in Chimera Checker tool (V). The same Hidden Markov Models (HMM)-based algorithm which was used to detect conserved regions (the end of rDNA small subunit, 5.8S, and the beginning of rDNA large subunit) inside fungal ITS sequences by ITS Extractor, was later used by Hartmann et al. (2010) for extracting hypervariable regions of bacterial, archaeal and fungal small subunit (16S/18S) rDNA sequences in a V-Extractor tool.

The software tools we have developed have gained attention also by the INSDC curators – program for detecting and reorientating reverse complement sequences (Nilsson et al. 2010B) is used by the NCBI team for checking new submissions of fungal DNA sequences. INSDC has recently shown interest in determining all misidentifications and chimeric sequences flagged in UNITE *INSDC* dataset to contact the original submitters of these data (Schoch C, personal communication). Since July 2011, UNITE sequences are hyperlinked in ENA if present in both databases.

Future directions for developing PlutoF workbench in the long run include the adaption of MVC architectural pattern and replacing the current implementation in PHP with an implementation in django, a web application framework written in Python programming language. In addition a paradigm change from the current procedural to a more object-oriented is in place. This will allow the independent development of domain logic and user interface, and make the concurrent development of different system modules by several programmers easier.

CONCLUSIONS

1. The UNITE database and its analysis tools were developed in a need of a good reference dataset to enable fast and reliable molecular identification of ectomycorrhizal fungi. The growth of the number of fungal ITS sequences, species and genera represented in UNITE during the previous 5 years, the usage of UNITE reference dataset in several “inhouse” analysis pipelines, current usage statistics for the UNITE homepage and the high number of citations it has received to date, indicates its necessity and usage by the fungal research community.
2. The current status of public gene repositories (INSDC) with regard to misidentified, chimeric and low quality sequences as well as the insufficiency of metadata present, clearly refers to the necessity for adding third-party annotations. Since this possibility is not yet implemented in INSDC, but the data is valuable and needed for future research, correcting and annotating of INSDC data needs to be done locally. NCBI is already using the software we developed for identifying reverse complementary sequences in their new submissions. They are also interested in chimeric and misidentified sequences flagged as such in the UNITE *INSDC* dataset which allows to hope that correcting or tagging misidentified and low quality sequences in INSDC will be possible in the future. Currently the ENA LinkOut system is the best solution for providing extra information available in curated databases such as UNITE for the interested user.
3. PlutoF cloud (servers, web-based workbench and the underlying database structure) was developed for storing, managing and analyzing the biodiversity data relevant to ecological and taxonomical research. Its initial usage mainly for managing specimen data in scientific collections in Estonia has changed in that today it is used by Estonian and international research community for a wide range of activities, such as managing scientific collections and ecological studies, keeping the Estonian Species Registry, annotating fungal INSDC sequences and using software tools for analyzing molecular sequence data.

REFERENCES

1. Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U. 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* 186 (2): 281–285.
2. Altschul SF, Madden DL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25 (17): 3389–3402.
3. Barcode of Life. <http://www.barcodeoflife.org/>
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2008. GenBank. *Nucleic Acids Research* 36 (Database Issue): D25–D30.
5. Bidartondo MI, Bruns TD, Blackwell M, et al. 2008. Preserving accuracy in GenBank. *Science* 319: 1616.
6. BioCASE (The Biological Collection Access Service for Europe). <http://www.biocase.org/>
7. Bridge PD, Roberts PJ, Spooner BM, Panchal G. 2003. On the unreliability of published DNA sequences. *New Phytologist* 160: 43–48.
8. Brock PM, Döring H, Bidartondo MI. 2009. How to know unknown fungi: the role of a herbarium. *New Phytologist* 181: 719–724.
9. Bueé M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 pyrosequencing analyses of forest soils reveal an unexpected high fungal diversity. *New Phytologist* 184: 449–456.
10. Caporaso JG, Kuczynski J, Stombaugh J et al. 2010. QIIME allows analysis of high-throughput community sequence data. *Nature Methods* 7: 335–336.
11. DDBJ (DNA Data Bank of Japan). <http://www.ddbj.nig.ac.jp/>
12. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology* 72 (7): 5069–5072.
13. Durling MB, Clemmensen KE, Stenlid J, Lindahl B. 2011. SCATA – An efficient bioinformatic pipeline for species identification and quantification after high-throughput sequencing of tagged amplicons. (submitted)
14. ENA (European Nucleotide Archive). <http://www.ebi.ac.uk/ena/>
15. EOL (Encyclopedia of Life). <http://www.eol.org/>
16. Final Report Of The OECD Megascience Forum Working Group On Biological Informatics. 1999. <http://www.oecd.org/dataoecd/24/32/2105199.pdf>
17. GBIF (Global Biodiversity Information Facility). <http://www.gbif.org/>
18. Guralnick RP, Hill AW, Lane M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10: 663–672.
19. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. 2010. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16 S/18 S) ribosomal RNA gene sequences. *Journal of Microbiological Methods* 83 (2): 250–253.
20. Horton TR, Bruns TD. 2001. The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. *Molecular Ecology* 10: 1855–1871.
21. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* 8: R143.

22. iTOL (Interactive Tree Of Life). <http://itol.embl.de/>
23. Jumpponen A, Jones KL. 2009. Massively parallel 454-sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* 184: 438–448.
24. Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vrålstad T, Ursing BM. 2005. UNITE: A database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* 166 (3): 1063–1068.
25. Ligginstoffer AS, Youssef NH, Couger MB, Elshahed MS. 2010. Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. *ISME Journal* 4: 1225–1235.
26. Ludwig W, Strunk O, Westram R et al. 2004. ARB: a software environment for sequence data. *Nucleic Acids Research* 32 (4): 1363 – 1371.
27. Nagy LG, Petkovits T, Kovács GM, Voigt K, Vágvölgyi C, Papp T. 2011. Where is the unseen fungal diversity hidden? A study of *Mortierella* reveals a large contribution of reference collections to the identification of fungal environmental sequences. *New Phytologist* 191: 789–794.
28. NCBI (National Center for Biotechnology Information). <http://www.ncbi.nlm.nih.gov/>
29. Nilsson RH, Kristiansson E, Ryberg M, Larsson K-H. 2005. Approaching the taxonomic affiliation of unidentified sequences in public databases – an example from the mycorrhizal fungi. *BMC Bioinformatics* 6: 178.
30. Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson K-H, Kõljalg U. 2006. Taxonomic reliability of DNA sequences in public sequences databases: A fungal perspective. *PLoS ONE* 1: e59.
31. Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K. 2010A. Rethinking taxon sampling in the light of environmental sequencing. *Cladistics* 26: 1–7.
32. Nilsson RH, Veldre V, Zheng W, Eckart M, Branco S, Hartmann M, Quince C, Godhe A, Bertrand Y, Alfredsson JF, Larsson K-H, Kõljalg U, Abarenkov K. 2010B. A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation. *Mycoscience* 52 (4): 278–282.
33. Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M. 2009. Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist* 184: 424–437.
34. Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* 188: 223–241.
35. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35: 7188–7196.
36. Pyle RL. 2003. Taxonomer: a relational data model for managing information relevant to taxonomic research. *Phyloinformatics* 1: 1–54.
37. Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System (<http://www.bargoddinglife.org>). *Molecular Ecology Notes* 7 (3): 355–364.

38. Rosling A, Cox F, Cruz-Martinez K, Ihrmark K, Grelet G-A, Lindahl BD, Menkis A, James TY. 2011. Archaeorhizomycetes: Unearthing an Ancient Class of Ubiquitous Soil Fungi. *Science* 333: 876–879.
39. Ryberg M, Nilsson RH, Kristiansson E, Töpel M, Jacobsson S, Larsson e. 2008. Mining metadata from unidentified ITS sequences in GenBank: A case study in *Inocybe* (*Basidiomycota*). *BMC Evolutionary Biology* 8: 50.
40. Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH. 2009. An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytologist* 181: 471–477.
41. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Opem-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* 75 (23): 7537–7541.
42. TDWG (Biodiversity Information Standards). <http://www.tdwg.org/>
43. Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U. 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* 188 (1): 291–301.
44. Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P. 2010. Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. *Research in Microbiology* 161: 439–445.
45. Vilgalys R. 2003. Taxonomic misidentification in public DNA databases. *New Phytologist* 160: 4–5.
46. Yilmaz P, Kottmann R, Field D, et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (X) sequence (MIxS) specifications. *Nature Biotechnology* 29 (5): 415–420.

SUMMARY IN ESTONIAN

PlutoF pilv – elurikkuse andmebaaside ja analüüsiplatvorm bioloogile

Elurikkuse informaatika tegeleb infotehnoloogiliste lahenduste loomise ja rakendamisega kogu elurikkust hõlmava andmestiku (ökoloogia, geneetika, taksonoomia jm) talletamiseks, sorteerimiseks ja analüüsimiseks. Elurikkuse informaatika arengu peamised suunad on:

1. automatiseeritud lahenduste loomine elurikkuse andmete kvaliteedi kontrolliks ja selle parandamiseks;
2. elurikkuse standardite loomine andmeväljade ühtlustamiseks, et andmed erinevates andmebaasides oleksid võrreldavad ja koos kasutatavad;
3. võimsamate otsingumootorite, andmete kuvamise ja analüüsimise tarkvara loomine.

Tänapäeval moodustab taksonite kohta käivast informatsioonist suure osa nende määramiseks kasutatav molekulaarne andmestik, sh. DNA nukleotiidsed järjestused. Seeneliikide määramisel keskkonnaproovidest (taimejuurtest, mullast, õhust, jm) on DNA põhine määramine möödapääsmatu, kuna seened esinevad neis proovides valdavalt seeneniidistikuna, mida ei ole võimalik morfoloogiliste või anatoomiliste tunnuste abil määrata. Kõige levinumaks meetodiks seente DNA põhisel määramisel on rDNA ITS järjestuste sarnasuse võrdlemine kasutades BLAST algoritmi. ITS järjestused on enamikus seenerühmades sõsarliikide eristamiseks piisavalt varieeruvad. Lisaks sellele võib neid genomis olla kuni mitusada koopiat. See teeb seeneliigi proovist määramise võimalikuks ka väga väikese rakkude arvu korral. BLAST algoritmi kasutamine eeldab aga referents DNA järjestuste andmebaasi olemasolu, mis 1) oleks piisava taksonoomilise katvusega, ning 2) sisaldaks eksperdi poolt määratud ja annoteeritud DNA järjestusi, mis pärinevad seene viljakehast või kultuurist. Juba 1990ndate algusest alates on suurimaks nukleotiidsete järjestuste andmebaasiks olnud avalike geenipankade konsortsium *International Nucleotide Sequence Database Collaboration* (INSDC), mis lisaks DNA järjestustele pakub ka erinevaid võimalusi BLAST algoritmi-põhisteks otsinguteks ning fülogeneesipuude joonistamiseks. Viimase kümne aasta jooksul läbi viidud uuringud on välja toonud mitmed INSDC andmebaaside kitsaskohad, mis takistavad neil olemast seente DNA põhisel määramisel referents andmebaasiks, nimelt: 1) kuigi taksonoomiliselt katvuselt kõige täielikum, sisaldab ta seente ITS järjestusi vähem kui 1% liikide kohta (seeneliikide hinnanguline koguarv on 1.5 miljonit); 2) küllatki suur osa INSDC andmebaasis olevatest ITS järjestustest on kehva kvaliteediga, valesti määratud või puuduliku metaandmestikuga. DNA järjestuste ja nendega seotud metaandmestiku kvaliteediprobleemide lahendamine nõuab: 1) kureeritava(te) referents andmebaasi(de) loomist; ning 2) INSDC andmestiku kontrollimist, parandamist ja täiendamist.

Käesoleva töö eesmärgiks oli luua: 1) seente DNA-põhiseks määramiseks vajalik referents andmebaas, mis sisaldaks kvaliteetseid, rikkaliku metaandmes-tikuga ja eksperdi poolt määratud seente viljakehadest eraldatud rDNA ITS järjestusi; 2) seente DNA-põhiseks määramiseks vajalikud tarkvaralahendused ja veebikeskkond nende kasutamiseks; 3) süsteem INSDC avalike seente rDNA ITS järjestuste kvaliteedi kontrollimiseks ja annoteringute lisamiseks; 4) e-taristu elurikkuse andmebaaside talletamiseks; ning 5) veebitöölaud elurikkuse andmestiku (sh Eesti liikide nimestik, eksemplarid teaduslikes kogudes, ökoloogilised uuringud ja nendega seotud molekulaarne andmestik, jm) sisestamiseks, haldamiseks ja analüüsimiseks.

Seente ITS järjestuste referents andmebaas realiseeriti UNITE andmebaasi loomisega 2003. aastal veebiaadressil <http://unite.zbi.ee> (hiljem juba aadressil <http://unite.ut.ee>). Algselt oli UNITE eesmärgiks Balti- ja Põhjamaade ektomükoriisat moodustavate seente referents andmebaasi loomine. Hiljem see piirang eemaldati ja andmebaas katab nüüdseks kogu seeneriiki. 2011. a. augustiks on andmebaasis olevate ITS referents järjestuste arv rohkem kui kolmekordistunud, ning unikaalsete seeneliikide ja -perekondade arv rohkem kui kahekordistunud. Aastast 2010 on UNITE andmebaas avatud ka keskkonna-proovidest pärit rikkaliku metaandmestikuga kvaliteetsetele DNA järjestustele, mida on võimalik eraldi andmestikuna analüüsidesse kaasata.

Seente DNA-põhiseks määramiseks tarviliku tarkvara arendamine algas 2003ndal aastal, kui UNITE avalikul kodulehel oli võimalik kasutada BLAST algoritmi-põhiseid sarnasuse otsingu programme (*blastn*, *galaxieBLAST*, *galaxieHMM*). Sellele järgnesid INSDC andmestiku kaasamine otsingutesse aastal 2006 ning *massBLASTer* programmi loomine aastal 2010. Viimane võimaldab BLAST algoritmi-põhisel määramisel analüüsida korraka tuhandeid DNA järjestusi. Lisaks kirjeldasime aastal 2010 tarkvara ITS järjestuste 2 erineva regiooni, so. ITS1 ja ITS2, äratundmiseks ja lõikamiseks, kimäärsete ITS järjestuste kindlakstegemiseks ning pürosekvenceerimise tulemusel saadud DNA järjestuste analüüsimiseks.

INSDC andmete annoteerimiseks loodi süsteem, kus kõik INSDC seente rDNA ITS järjestused laetakse perioodiliselt alla PlutoF pilve andmebaasi. See võimaldab nende järjestuste kvaliteedi kontrollimist, määrangute lisamist ja metaandmestiku täiendamist. Selline annoteeritud andmestik on suureks abiks üha kasvavate molekulaarsete andmemassiivide analüüsimisel nii seenekoos-luste kindlakstegemisel erinevates keskkonnaproovides kui ka teistes ökoloogilistes, biogeograafilistes ja fülogeneesi uuringutes.

PlutoF pilves asuvate kümnete andmebaaside (UNITE, INSDC, geenijärjes-tuste annoteringud, tööruhmade ja üksikisikute andmestik) haldamiseks loodi veebi-töölaud, mis võimaldab kasutajal sisestada ja hallata elurikkuse andmeid taksonoomiast keskkonnagenoomikani ning kasutada molekulaarsete andmete analüüsimiseks eelpool loetletud tarkvara. PlutoF pilve veebi-töölaud võimaldab samaaegselt hallata nii oma isiklikke kui ka tööruhma või projekti andmebaase, soodustades sellega erinevate rahvusvaheliste tööruhmade koostööd. PlutoF pilv on e-taristu, mille moodustavad serverid, veebi-töölaud ja andme-

baasides hoitav andmestik (vt. Joonis 1). PlutoF pilvest pärinevad mitmed avalikud veebiväljundid, sh. UNITE (<http://unite.ut.ee>) ja eElurikkus (<http://elurikkus.ut.ee>). PlutoF pilve on edukalt rakendatud seente ökoloogia ja taksonoomia alase teadustöö tegemiseks aga ka loodusteaduslike kogude andmebaasistamiseks ja Eesti Liikide Registri koostamiseks.

ACKNOWLEDGEMENTS

I am very grateful to Urmas Kõljalg for becoming my supervisor for almost 10 years ago and guiding me through these years full of optimism and enthusiasm to our research.

I am particularly grateful to Henrik Nilsson and Leho Tedersoo for all their help, support and time.

I want to thank all my coworkers (without institutional boundaries) and co-authors with special thanks to Triin Naadel, Heidi Tamm, Teele Jairus and Jane Oja for their fruitful discussions during lunch times and long walks in the forest in mycology camps, and Irja Saar, Kadri Põldmaa and Erast Parmasto for their help and fresh ideas to my work. You have all greatly contributed to this thesis.

My special gratitude goes to all my friends, especially to Natalja, the girl with all the questions, and Indrek, the boy with all the answers in the world.

I also thank my parents and brothers for always being there when I needed you. You're the best!

My activities received support from ESF grants no 8235 and 6606, FIBIR, Doctoral School of Ecology and Environmental Sciences and Doctoral School of Earth Sciences and Ecology.

The development of species registry was supported in 2008–2010 by grant EE0018 “Estonian Biodiversity data base and information network supporting Natura 2000”.

Supplementary Item 1.
Species information page for *Thelephora terrestris* Ehrh. on Estonian eBiodiversity.

eBiodiversity [Front page](#) [Species](#) [Collections](#) [Observations](#) [Dictionary](#) [Workgroups](#) [News](#) [Quick search](#) [In English](#)

[Overview](#)

Classification


[Quick search](#)

[Advanced search](#)


[Estonian Red List of Threatened Species](#)

Thelephora terrestris Ehrh.
Harilik lehtmähkis

- [Placement in taxonomic system](#)
- [Database records on the map](#)
- [Reference data about species occurrence in Estonia](#)
- [Specimens in scientific collections](#)
- [Species observations](#)
- [Public gene sequences](#)



Author: Umas Kõljalg, specimen: TU100201



Synonyms:
Thelephora laciniata (Pers.) Pers.


Placement in taxonomic system
* only species found in Estonia are shown

Species placement in taxonomic system:
Fungi; Basidiomycota; Agaricomycotina; Agaricomycetes; Incertae sedis; Thelephorales; Thelephoraceae; Thelephora; Thelephora terrestris

Your location on the tree:

- kgl Fungi; Seened
- phy Basidiomycota; Kandseened
- stp Agaricomycotina
- cks Agaricomycetes
- sbc Incertae sedis
- ord Thelephorales; Lehtmähkiseladised
- fam Thelephoraceae; Lehtmähkiselised
- gen Thelephora; Lehtmähkis
- spe Thelephora terrestris; Harilik lehtmähkis

Databased records on the map



Map | Satellite | Hybrid

Legend:

- Specimen in collection
- Photo in database
- Species observation
- Photo in database
- Reference in database
- Specimen with gene sequence in database
- Specimen with photo and gene sequence in database
- Gene sequences from environmental studies

Reference data about species occurrence in Estonia

1. Järva L., Parmasto E. 1980. Eesti seenite koondnimeistik. Lk. 134
Inserted by: Erast Parmasto, 30.05.2008
2. Järva L., Parmasto I., Vaasma M. 1998. Eesti seenite koondnimeistik. 1. liidenduskõide (1975-1990). Lk. 82
Inserted by: Erast Parmasto, 04.08.2009
3. Keizer G.J. 2005. Seente entsüklopeedia. Lk. 101
Inserted by: Erast Parmasto, 03.02.2011
4. Kõljalg U., Dahlberg A., Taylor A.F.S., Larsson E., Hallenberg N., Stenlid J., Larsson K.-H., Fransson P.M., Karen O., Jonsson L. 2000. Diversity and abundance of resupinate telephoroid fungi as ectomycorrhizal symbionts in Swedish boreal forests.
Inserted by: Urmas Kõljalg, 12.09.2002
5. Parmasto E., Kalamees K., Kaimel U., Parmasto I., Raatvir A., Vaasma M. 2004. Järvsejja kattealuse põllumetsa seenestik. Lk. 132
Inserted by: Erast Parmasto, 21.09.2009
6. Parmasto E., Parmasto I. 2006. Fungi of Ruhnu Island (Estonia). Estonia Maritime, 7. Lk. 55
Inserted by: Erast Parmasto, 19.08.2009
7. Tederso L., Suvil T., Jalrus T., Kõljalg U. 2007. Forest microsite effects on community composition of ectomycorrhizal fungi on seedlings of Picea abies and Betula pendula
Inserted by: Katrin Kohe, 24.10.2008

Thelephora laciniata (Pers.) Pers.

Currently there are no references linked to this species in our database.

Specimens in scientific collections

EAA013393; EAA013394; EAA013395; EAA013396; EAA013398; EAA013399; EAA013400; EAA013401; EAA013402; EAA013403; EAA013404; EAA013405; EAA013406; TAAM105210 (Det.: Urmas Kõljalg); TAAM108027 (Det.: Urmas Kõljalg); TAAM127701; TAAM128270; TAAM153687 (Det.: Erast Parmasto); TAAM153710 (Det.: Erast Parmasto); TAAM159495 (Det.: Urmas Kõljalg)
(20 first of 49 are shown)

[View all specimens of that species in the National database of Estonian fungal collections](#)

Thelephora laciniata (Pers.) Pers.

TAAM153681; TAAM184934 (Det.: Erast Parmasto)

[View all specimens of that species in the National database of Estonian fungal collections](#)

Specimens without geographical coordinates

EAA013396; EAA013396; EAA013398; EAA013399; EAA013400; EAA013401; EAA013402; EAA013403; EAA013404; EAA013405; EAA013406; EAA013394; EAA013393; TAAM105210; TAAM127701; TAAM184933; TAAM108027; TU108366; TU108604; TALL F000004

Species observations

13.09.2000 Põlva Co., Orava Comm., 8 km W of Orava. Vaatleja(d): Erast Parmasto
04.09.2000 Hiium Co., Kõrgessaare Comm., Linnaru Nature Reserve. Vaatleja(d): Erast Parmasto
09.04.2000 Hiium Co., Kõrgessaare Comm., Kõpu. Vaatleja(s): Erast Parmasto
29.07.2000 Võru Co., Antsla Comm., Roosiku E of Tsooru. Vaatleja(s): Erast Parmasto

[Vaata kõiki selle liigi vaatlusi andmebaasis](#)

Thelephora laciniata (Pers.) Pers.

Currently there are no observations of this species in our database.

Public gene sequences

* only gene sequences originating from specimens collected in Estonia are shown

UNITE:UD800087 | UK035 (ITS1 - 5.8S - ITS2)
UNITE:UD8000215 | [GenBank: AF272923](#) | UK14 (ITS1 - 5.8S - ITS2)
UNITE:UD8003346 | 334 (ITS1 - 5.8S - ITS2)
UNITE:UD8003345 | 808 (ITS1 - 5.8S - ITS2)
UNITE:UD8000224 | UK57 (ITS1 - 5.8S - ITS2)

Thelephora laciniata (Pers.) Pers.

Currently there are no gene sequences for this species in our database.

Latest news

11.11. United Nations Decade on Biodiversity
11.11. Species of the Day

Contact

E: cs@biocenter.ut.ee
A: Ravila 14A, 50411 Tartu



PUBLICATIONS

CURRICULUM VITAE

I. General

Name: Kessy Abarenkov
Date and place of birth: 18.12.1980, Tartu, Estonia
Citizenship and nationality: Estonian
Language skills: Estonian (mother tongue), English, Russian
Contact information: Institute of Ecology and Earth Sciences,
University of Tartu. 14A Ravila Street 50411
Tartu, Estonia;
Phone: +372 737 6175;
e-mail: kessy.abarenkov@ut.ee
Current position: University of Tartu, Natural History Museum,
information technology specialist

Education

1987–1999 Kivilinna Gymnasium of Tartu
1999–2004 University of Tartu, Bioinformatics, B.Sc.
2004–2006 University of Tartu, Botany and Mycology, M.Sc.
2006–2011 University of Tartu, PhD student in botany and mycology

II. Scientific and research activity

Main research interests

Biodiversity informatics tools and databases

Publications (CC)

Nilsson RH, Tedersoo L, Lindahl BD, Kjølner R, Carlsen T, Quince C, **Abarenkov K**, Pennanen T, Stenlid J, Bruns T, Larsson K-H, Kõljalg U, Kauserud H. 2011. Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. *New Phytologist* 191: 314–318.

Hartmann M, Howes CG, Veldre V, Schneider S, Vaishampayan PA, Yannarell AC, Quince C, Johansson P, Björkroth KJ, **Abarenkov K**, Hallam SJ, Mohn WW, Nilsson RH. 2011. V-RevComp: Automated high-throughput detection of reverse complementary 16S rRNA gene sequences in large environmental and taxonomic datasets. *FEMS Microbiology Letters* 319: 140–145.

Nilsson RH, Veldre V, Wang Z, Eckart M, Branco S, Hartmann M, Quince C, Godhe A, Bertrand Y, Alfredsson JF, Larsson K-H, Kõljalg U, **Abarenkov K**. 2011. A note on the incidence of reverse complementary fungal ITS

- sequences in the public sequence databases and a software tool for their detection and reorientation. *Mycoscience* 52 (4): 278–282.
- Nilsson RH, Ryberg M, Sjökvist E, **Abarenkov K**. 2011. Rethinking taxon sampling in the light of environmental sequencing. *Cladistics* 27: 197–203.
- Abarenkov K**, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Proust M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U. 2010. PlutoF – a web-based workbench for ecological and taxonomical research, with an on-line implementation for fungal ITS sequences. *Evolutionary Bioinformatics* 6: 189–196.
- Hartmann M, Howes CG, **Abarenkov K**, Mohn WW, Nilsson RH. 2010. V-Extractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods* 83: 250–253.
- Tedersoo L, Nilsson RH, **Abarenkov K**, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U. 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* 188 (1): 291–301.
- Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, **Abarenkov K**. 2010. An open source software package for rapid, automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology* 3: 284–287.
- Nilsson RH, **Abarenkov K**, Veldre V, Nylinder S, De Wit P, Brosche S, Alfredsson JF, Ryberg M, Kristianson E. 2010. An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources* 10: 1076–1081.
- Abarenkov K**, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U. 2009. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* 186 (2): 281–285.
- Suvi T, Tedersoo L, **Abarenkov K**, Beaver K, Gerlach J, Kõljalg U. 2009. Mycorrhizal symbionts of *Pisonia grandis* and *P. sechellarum* in Seychelles: identification of mycorrhizal fungi and description of new *Tomentella* species. *Mycologia* 102 (3): 522–533.
- Nilsson RH, Ryberg M, **Abarenkov K**, Sjökvist E, Kristiansson E. 2009. The ITS region as target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters* 296(1): 97–101.
- Tedersoo L, Jairus T, Horton BM, **Abarenkov K**, Suvi T, Saar I, Kõljalg U. 2008. Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. *New Phytologist* 180 (2): 479–490.

- Nilsson RH, Ryberg M, Kristiansson E, **Abarenkov K**, Larsson K-H, Kõljalg U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS ONE 1: e59
- Kõljalg U, Larsson K-H, **Abarenkov K**, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjöller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist 166: 1063–1068.

Conference presentations

- Abarenkov K**, Kõljalg U, Nilsson RH, Parmasto E, Tedersoo L, Kuslapuu A. PlutoF – the Web-based Workbench for Managing Data in the UNITE Database. IMC9: The Biology of Fungi. August 2010, Edinburgh, United Kingdom.
- Kõljalg U, **Abarenkov K**, Kuslapuu A, Parmasto E, Kurina O, Luig J, Martin M, Vellak K, Aan A, Reier Ü, Timm T, Kukk T, Orav K, Saar I, Pärtel K, Proust M, Õunap E, Timm U, Puura I. Estonian Biodiversity Database. The e-Biosphere 09 International Conference on Biodiversity Informatics. June 2009, London, United Kingdom.
- Abarenkov K**, Kõljalg U, Suvi T, Jairus T, Saar I, Tedersoo L. High global diversity of *Tomentella* and *Thelephora* (Thelephorales, Basidiomycota). 21st New Phytologist Symposium – The Ecology of Ectomycorrhizal Fungi. December 2008, Montpellier, France.

Scholarships

- 2010 Doctoral School of Earth Sciences and Ecology PhD student scholarship
- 2009 Doctoral School of Ecology and Environmental Sciences PhD student scholarship
- 2008 Doctoral School of Ecology and Environmental Sciences PhD student scholarship

Participation in international courses

- Practical course “Molecular Evolution, Phylogenetics and Adaption”. February 2010, Portugal.
- Practical course “Computational phyloinformatics”. July 2009, Portugal.
- PhD course “Identifying ectomycorrhizal fungi – from environmental samples to DNA sequences”. August 2007, Denmark.
- PhD summer school “PhD Summer School in Biodiversity Informatics”. August 2006, Denmark.

CURRICULUM VITAE

I. Üldandmed

Ees- ja perekonnanimi: Kessy Abarenkov
Sünniaeg ja -koht: 18.12.1980, Tartu, Eesti
Kodakondsus: Eesti
Keelteoskus: eesti, inglise, vene
Kontaktandmed: Tartu Ülikool, Ökoloogia ja Maateaduste Instituut.
Ravila 14A, 50411 Tartu, Eesti;
Tel.: +372 737 6175;
e-mail: kessy.abarenkov@ut.ee
Praegune töökoht: Tartu Ülikooli Loodusmuuseum, infotehnoloogia spetsialist

Haridus

1987–1999 Tartu Kivilinna Gümnaasium
1999–2004 Tartu Ülikool, B.Sc. bioinformaatika erialal
2004–2006 Tartu Ülikool, M.Sc. botaanika ja mükoloogia erialal
2006–2011 Tartu Ülikool, doktorant botaanika ja mükoloogia erialal

II. Teaduslik ja arendustegevus

Peamised uurimisvaldkonnad

Elurikkuse informaatika rakendused ja andmebaasid

Teaduspublikatsioonide loetelu

Nilsson RH, Tedersoo L, Lindahl BD, Kjølner R, Carlsen T, Quince C, **Abarenkov K**, Pennanen T, Stenlid J, Bruns T, Larsson K-H, Kõljalg U, Kauserud H. 2011. Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. *New Phytologist* 191: 314–318.

Hartmann M, Howes CG, Veldre V, Schneider S, Vaishampayan PA, Yannarell AC, Quince C, Johansson P, Björkroth KJ, **Abarenkov K**, Hallam SJ, Mohn WW, Nilsson RH. 2011. V-RevComp: Automated high-throughput detection of reverse complementary 16S rRNA gene sequences in large environmental and taxonomic datasets. *FEMS Microbiology Letters* 319: 140–145.

Nilsson RH, Veldre V, Wang Z, Eckart M, Branco S, Hartmann M, Quince C, Godhe A, Bertrand Y, Alfredsson JF, Larsson K-H, Kõljalg U, **Abarenkov K**. 2011. A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation. *Mycoscience* 52 (4): 278–282.

- Nilsson RH, Ryberg M, Sjökvist E, **Abarenkov K**. 2011. Rethinking taxon sampling in the light of environmental sequencing. *Cladistics* 27: 197–203.
- Abarenkov K**, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Proust M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U. 2010. PlutoF – a web-based workbench for ecological and taxonomical research, with an on-line implementation for fungal ITS sequences. *Evolutionary Bioinformatics* 6: 189–196.
- Hartmann M, Howes CG, **Abarenkov K**, Mohn WW, Nilsson RH. 2010. V-Extractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods* 83: 250–253.
- Tedersoo L, Nilsson RH, **Abarenkov K**, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U. 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* 188 (1): 291–301.
- Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, **Abarenkov K**. 2010. An open source software package for rapid, automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology* 3: 284–287.
- Nilsson RH, **Abarenkov K**, Veldre V, Nylinder S, De Wit P, Brosche S, Alfredsson JF, Ryberg M, Kristianson E. 2010. An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources* 10: 1076–1081.
- Abarenkov K**, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U. 2009. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* 186 (2): 281–285.
- Suvi T, Tedersoo L, **Abarenkov K**, Beaver K, Gerlach J, Kõljalg U. 2009. Mycorrhizal symbionts of *Pisonia grandis* and *P. sechellarum* in Seychelles: identification of mycorrhizal fungi and description of new *Tomentella* species. *Mycologia* 102 (3): 522–533.
- Nilsson RH, Ryberg M, **Abarenkov K**, Sjökvist E, Kristiansson E. 2009. The ITS region as target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters* 296(1): 97–101.
- Tedersoo L, Jairus T, Horton BM, **Abarenkov K**, Suvi T, Saar I, Kõljalg U. 2008. Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. *New Phytologist* 180 (2): 479–490.
- Nilsson RH, Ryberg M, Kristiansson E, **Abarenkov K**, Larsson K-H, Kõljalg U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE* 1: e59

Kõljalg U, Larsson K-H, **Abarenkov K**, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* 166: 1063–1068.

Konverentsi tekkanded

Abarenkov K, Kõljalg U, Nilsson RH, Parmasto E, Tedersoo L, Kuslapuu A. PlutoF – the Web-based Workbench for Managing Data in the UNITE Database. IMC9: The Biology of Fungi. August 2010, Edinburgh, Suurbritannia.

Kõljalg U, **Abarenkov K**, Kuslapuu A, Parmasto E, Kurina O, Luig J, Martin M, Vellak K, Aan A, Reier Ü, Timm T, Kukk T, Orav K, Saar I, Pärtel K, Prou M, Õunap E, Timm U, Puura I. Estonian Biodiversity Database. The e-Biosphere 09 International Conference on Biodiversity Informatics. Juuni 2009, London, Suurbritannia.

Abarenkov K, Kõljalg U, Suvi T, Jairus T, Saar I, Tedersoo L. High global diversity of *Tomentella* and *Thelephora* (Thelephorales, Basidiomycota). 21st New Phytologist Symposium – The Ecology of Ectomycorrhizal Fungi. Detsember 2008, Montpellier, Prantsusmaa.

Saadud uurimistoetused

- | | |
|------|---|
| 2010 | Maateaduste ja ökoloogia doktorikool, toetus doktorandi teadustöö finantseerimiseks |
| 2009 | Ökoloogia ja keskkonnateaduste doktorikool, toetus doktorandi teadustöö finantseerimiseks |
| 2008 | Ökoloogia ja keskkonnateaduste doktorikool, toetus doktorandi teadustöö finantseerimiseks |

Erialane enesetäiendus

- Osavõtt praktilisest kursusest “Molecular Evolution, Phylogenetics and Adaption”, veebruaris 2010 Portugalis.
- Osavõtt praktilisest kursusest “Computational phyloinformatics”, juulis 2009 Portugalis.
- Osavõtt doktorantide koolist “Identifying ectomycorrhizal fungi – from environmental samples to DNA sequences”, augustis 2007 Taanis.
- Osavõtt doktorantide suvekoolist “PhD Summer School in Biodiversity Informatics”, augustis 2006 Taanis.

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 Käär.** The development of an automatic online dynamic fluorescence-based pH-dependent fiber optic penicillin flowthrough 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 environmental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben.** Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes.** Respiration rhythms 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 temperate 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 crassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina.** Fungus gnats in Estonia (*Diptera: Bolitophilidae, Kero-platidae, 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 erotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O₃ and CO₂ on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptone-mal 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 popula-tions: 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 invasions 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 Northern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn4652 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 N₂ fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.

102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.
103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.
105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.
106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.
109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn4652 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 Tn4652 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 CO₂ 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, 2008, 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 Säälük.** 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 Tn4652 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 Väriv.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Väik.** 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. **Teale Jairus.** Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.