SQUEEZING MORE INFORMATION OUT OF BIOLOGICAL DATA - DEVELOPMENT AND APPLICATION OF BIOINFORMATIC TOOLS FOR ECOLOGY, EVOLUTION AND GENOMICS

MEHR AUS BIOLOGISCHEN DATEN HERAUSHOLEN - ENTWICKLUNG UND ANWENDUNG BIOINFORMATISCHER PROGRAMME FÜR ÖKOLOGIE, EVOLUTION UND GENOMIK

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-University of Würzburg,
Section: Integrative Biology

submitted by
MARKUS JOHANNES ANKENBRAND

from
Würzburg

Würzburg 2017

SUBMITTED ON: ....

Office stamp

MEMBERS OF THE *PROMOTIONSKOMITEE*:

Chairperson:	*Prof. Dr. Thomas Schmitt*
Primary Supervisor:	*PD Dr. Alexander Keller*
Supervisor (Second):	*Prof. Dr. Jörg Schultz*
Supervisor (Third):	*Dr. Robert R. Junker*

DATE OF PUBLIC DEFENCE: ....

DATE OF RECEIPT OF CERTIFICATES: ....

SUMMARY

New experimental methods have drastically accelerated the pace and quantity at which biological data is generated. High-throughput DNA sequencing is one of the pivotal new technologies. It offers a number of novel applications in various fields of biology, including ecology, evolution, and genomics. However, together with those opportunities many new challenges arise. Specialized algorithms and software are required to cope with the amount of data, often requiring substantial training in bioinformatic methods. Another way to make those data accessible to non-bioinformaticians is the development of programs with intuitive user interfaces.

In my thesis I developed analyses and programs to tackle current problems with high-throughput data in biology. In the field of ecology this covers the establishment of the bioinformatic workflow for pollen DNA meta-barcoding. Furthermore, I developed an application that facilitates the analysis of ecological communities in the context of their traits. Information from multiple public databases have been aggregated and can now be mapped automatically to existing community tables for interactive inspection. In evolution the new data are used to reconstruct phylogenetic trees from multiple genes. I developed the tool bcgTree to automate this process for bacteria. Many plant genomes have been sequenced in current years. Sequencing reads of those projects also contain data from the chloroplasts. The tool chloroExtractor supports the targeted extraction and analysis of the chloroplast genome. To compare the structure of multiple genomes specialized software is required for calculation and visualization of the relationships. I developed AliTV to address this. In contrast to existing programs for this task it allows interactive adjustments of produced graphics. Thus facilitating the discovery of biologically relevant information. Another application I developed helps to analyze transcriptomes even if no reference genome is present. This is achieved by aggregating the different pieces of information, like functional annotation and expression level, for each transcript in a web platform. Scientists can then search, filter, subset, and visualize the transcriptome.

Together the methods and tools expedite insights into biological systems that were not possible before.

ZUSAMMENFASSUNG

Neue experimentelle Methoden haben die Geschwindigkeit und Masse, in der biologische Daten generiert werden, in den letzten Jahren enorm gesteigert. Eine zentrale neue Technologie ist die Hochdurchsatzsequenzierung von DNA. Diese Technik eröffnet eine ganze Reihe Anwendungsmöglichkeiten in vielen Bereichen der Biologie, einschließlich der Ökologie, Evolution und Genomik. Neben den neuen Möglichkeiten treten jedoch auch neue Herausforderungen auf. So bedarf es spezialisierter Algorithmen und Computerprogramme, um mit der Masse an Daten umgehen zu können. Diese erfordern in der Regel ein fundiertes Training in bioinformatischen Methoden. Ein Weg, die Daten auch Wissenschaftlern ohne diesen Hintergrund zugänglich zu machen ist die Entwicklung von Programmen, die sich intuitiv bedienen lassen.

In meiner Doktorarbeit habe ich Analysen und Programme entwickelt, um einige aktuelle Probleme mit Hochdurchsatzdaten in der Biologie zu lösen. Im Bereich der Ökologie umfasst das die Etablierung der bioinformatischen Methode, um Pollen DNA Metabarcoding durchzuführen. Darüberhinaus habe ich eine Anwendung entwickelt, die es ermöglicht Artgemeinschaften im Kontext ihrer Eigenschaften zu erforschen. Dazu wurden Informationen aus diversen öffentlichen Datenbanken zusammen getragen. Diese können nun automatisch auf bestehende Projekte übertragen und interaktiv analysiert werden. Im Bereich der Evolution ermöglichen die neuen Daten phylogenetische Berechnungen mit multiplen Genen durchzuführen. Um dies für Bakterien zu automatisieren habe ich das Programm bcgTree entwickelt. In den letzten Jahren wurden viele pflanzliche Genome sequenziert. Die Sequenzdaten des pflanzlichen Genoms enthalten auch die des Chloroplasten. Das Programm chloroExtractor unterstützt die gezielte Analyse des Chloroplasten Genoms. Um jedoch die Struktur mehrerer Genome miteinander vergleichen zu können, wird spezielle Software benötigt, die den Vergleich berechnen und visuell darstellen kann. Daher habe ich das Programm AliTV entwickelt. Im Gegensatz zu bestehenden Programmen erlaubt AliTV interaktive Anpassungen der erzeugten Grafik. Das erleichtert es die relevanten Informationen zu finden. Ein weiteres von mir entwickeltes Programm hilft dabei Transkriptom Daten zu analysieren, auch wenn kein Referenzgenom vorliegt. Dazu werden Informationen zu jedem Transkript, z.B. Funktion und Expressionslevel, in einer Webanwendung aggregiert. Forscher können diese durchsuchen, filtern und graphisch darstellen.

Zusammen eröffnen die entwickelten Methoden und Programme die Möglichkeit, Erkenntnisse über biologische Systeme zu erlangen, die bislang nicht möglich waren.

PUBLICATIONS

This is the list of publications I co-authored related to the topic of my PhD thesis. First author publications are indicated by a ★ symbol in the margin.

Original research publications and a preprint included in this PhD thesis (note, that publication 4. is accepted but not yet published, while 9. has just been submitted and is not yet reviewed):

1. Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F. 2015. "ITS2 Database V: Twice as Much." *Molecular Biology and Evolution* 32(11):3030–3032.
   DOI: 10.1093/molbev/msv174 ★

2. Keller A, Danner N, Grimmer G, Ankenbrand M, von der Ohe K, von der Ohe W, Rost S, Härtel S, Steffan-Dewenter I. 2015. "Evaluating Multiplexed Next-Generation Sequencing as a Method in Palynology for Mixed Pollen Samples." *Plant Biology* 17(2):558–566.
   DOI:10.1111/plb.12251

3. Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I, Keller A. 2015. "Increased Efficiency in Identifying Mixed Pollen Samples by Meta-Barcoding with a Dual-Indexing Approach." *BMC Ecology* 15(1):20.
   DOI:10.1186/s12898-015-0051-y

4. Sickel W*, Ankenbrand MJ*, Grimmer G, Förster F, Steffan-Dewenter I, Keller A. (in press) "Standard Method for Identification of Bee Pollen Mixtures Through Meta-Barcoding." *COLOSS BEEBOOK Volume III* ★

5. Keller A, Grimmer G, Sickel W, Ankenbrand MJ. 2016. "DNA-Metabarcoding – ein neuer Blick auf organismische Diversität." *BIOspektrum* 22(2):147–150.
   DOI:10.1007/s12268-016-0669-0

6. Ankenbrand MJ, Terhoeven N, Hohlfeld S, Förster F, Keller A. 2017. "Biojs-Io-Biom, a BioJS Component for Handling Data in Biological Observation Matrix (BIOM) Format." *F1000Research* 5:2348.
   DOI:10.12688/f1000research.9618.2 ★

7. Ankenbrand MJ, Hohlfeld SCY, Förster F, Keller A. 2017. "FENNEC - Functional Exploration of Natural Networks and Ecological Communities." *bioRχiv* (preprint, not peer reviewed)
   DOI:10.1101/194308 ★

★ 8. Ankenbrand MJ, Keller A. 2016. "BcgTree: Automatized Phylogenetic Tree Building from Bacterial Core Genomes." *Genome* 59(10):783-791
doi:10.1139/gen-2015-0175

★ 9. Ankenbrand MJ*, Pfaff S*, Terhoeven N, Qureischi M, Gündel M, Weiß CL, Hackl T, Förster F. (submitted) "ChloroExtractor: Extraction and Assembly of the Chloroplast Genome from Whole Genome Shotgun Data" *The Journal of Open Source Software* (not yet peer reviewed)

★ 10. Ankenbrand MJ*, Hohlfeld S*, Hackl T, and Förster F. 2017. "AliTV – Interactive Visualization of Whole Genome Comparisons." *PeerJ Computer Science* 3:e116.
doi:10.7717/peerj-cs.116

★ 11. Ankenbrand MJ*, Weber L*, Becker D, Förster F, Bemm F. 2016. "TBro: Visualization and Management of de Novo Transcriptomes." *Database* 2016(0):baw146.
doi: 10.1093/database/baw146

Original research publications outside the scope of this PhD thesis:

12. Bemm F, Becker D, Larisch C, Kreuzer I, Escalante-Perez M, Schulze WX, Ankenbrand M, Van de Weyer AL, Krol E, Al-Rasheid KA, Mithöfer A, Weber AP, Schultz J, Hedrich R. 2016. "Venus Flytrap Carnivorous Lifestyle Builds on Herbivore Defense Strategies." *Genome Research* 26(6):812-825.
doi:10.1101/gr.202200.115.

13. Wolf M, Chen S, Song J, Ankenbrand M, Müller T. 2013. "Compensatory Base Changes in ITS2 Secondary Structures Correlate with the Biological Species Concept Despite Intragenomic Variability in ITS2 Sequences – A Proof of Concept." *PLoS ONE* 8(6):e66726.
doi: 10.1371/journal.pone.0066726

* These authors contributed equally.

Any technology distinguishable from magic is insufficiently advanced.

— Gehm's corollary to Clarke's third law

ACKNOWLEDGMENTS

It has been a long chain of events leading to this point: me trying to find the words to thank all the wonderful people that helped me along the way. I'll try to trace this path back in reverse chronological order. The last three and a half years I actually worked on this thesis which consists of multiple projects with many nice people. Foremost, there is my primary supervisor PD Dr. Alexander Keller, I can not express how grateful I am for all the support you gave me. You created the perfect environment for me to grow, lots of freedom to develop my own ideas but always helpful advice and guidance when I needed it. Also a big thanks to my other thesis supervisors Prof. Dr. Jörg Schultz and Dr. Robert Junker for your valuable feedback and support. Dr. Frank Förster you have been a mentor and friend throughout my whole academic career. I thank the people from the Center of Computational and Theoretical Biology as well as those from the departments of Bioinformatics, and Animal Ecology & Tropical Biology. Specifically my group, office, and GSLS mates Dr. Wiebke Sickel, Sonja Hohlfeld, Niklas Terhoeven, Torsten Paul, Jan Freudenthal, Ludwig Leidinger, Anna Voulgari Kokota, Prof. Philip Kollmannsberger, Prof. Arthur Korte, Manuela Engert, and Franziska Saul. I am grateful to the Graduate School of Life Sciences for the scholarship and for everyone of the staff being ever supportive and helpful. All the credit for the LaTeXstyle of this thesis belongs to Andreas Leimbach and Dr. Frank Förster. Past supervisors did a great job in keeping my passion for science alive, in particular Dr. Thomas Hackl, Felix Bemm, Lenz Weber, Dr. Matthias Wolf, and Dr. Tobias Müller. I also won't fail to mention two outstanding school teachers that managed to kindle my interest in science, Martin Feige and Christian Lorey. Of course I would not have been able to accomplish anything if not for my family and friends. I can not find the words to properly express my feelings. Thanks to my parents Rainer and Susanne for giving me everything. Last but foremost Hannah for your ∞∞ love.

CONTENTS

I INTRODUCTION 1

1 INCREASING VOLUME OF DATA IN BIOLOGY 3

- 1.1 The Bioinformatic Bottleneck 3
- 1.2 Challenges of Research Software Engineering 3

2 APPLICATIONS IN ECOLOGY, EVOLUTION, AND GENOMICS 5

- 2.1 Ecology 5
- 2.2 Evolution 8
- 2.3 Genomics 9

3 AIM OF THIS THESIS 13

II MATERIAL AND METHODS 15

4 SOURCE CODE 17

- 4.1 meta-barcoding-dual-indexing 17
- 4.2 biojs-io-biom 18
- 4.3 biom-conversion-server 18
- 4.4 FENNEC 19
- 4.5 bcgTree 20
- 4.6 chloroExtractor 20
- 4.7 AliTV 21
- 4.8 TBro 21

III RESULTS 23

5 PUBLICATIONS 25

- 5.1 ITS2 Database V: Twice as Much 26
- 5.2 Evaluating Multiplexed Next-Generation Sequencing as a Method in Palynology for Mixed Pollen Samples . . . 38
- 5.3 Increased Efficiency in Identifying Mixed Pollen Samples by Meta-Barcoding with a Dual-Indexing Approach 48
- 5.4 Standard Method for Identification of Bee Pollen Mixtures Through Meta-Barcoding 70
- 5.5 DNA-Metabarcoding - ein neuer Blick auf organismische Diversität 82
- 5.6 biojs-io-biom, a BioJS Component for Handling Data in Biological Observation Matrix (BIOM) Format 86
- 5.7 FENNEC - Functional Exploration of Natural Networks and Ecological Communities 101
- 5.8 bcgTree: Automatized Phylogenetic Tree Building from Bacterial Core Genomes 122
- 5.9 chloroExtractor: Extraction and Assembly of the Chloroplast Genome from Whole Genome Shotgun Data . . . 135
- 5.10 AliTV-Interactive Visualization of Whole Genome Comparisons 139

5.11 TBro: Visualization and Management of de Novo Transcriptomes . . . 156

IV DISCUSSION 167

6 GETTING MORE OUT OF BIOLOGICAL DATA 169

7 ADVANCES IN ECOLOGY, EVOLUTION, AND GENOMICS 171

7.1 Ecology . . . 171

7.2 Evolution . . . 172

7.3 Genomics . . . 174

8 FUTURE WORK 177

9 CONCLUSION 179

V APPENDIX 181

A INDIVIDUAL AUTHOR CONTRIBUTIONS 183

B CURRICULUM VITÆ 195

BIBLIOGRAPHY 197

LIST OF TABLES

Table 1 Individual author contributions for each part of Ankenbrand et al. (2015) 183
Table 2 Individual author contributions for the figures and tables of Ankenbrand et al. (2015) 183
Table 3 Individual author contributions for each part of Keller et al. (2015) 184
Table 4 Individual author contributions for the figures and tables of Keller et al. (2015) 184
Table 5 Individual author contributions for each part of Sickel et al. (2015) 185
Table 6 Individual author contributions for the figures and tables of Sickel et al. (2015) 185
Table 7 Individual author contributions for each part of Sickel et al. (in press) 186
Table 8 Individual author contributions for the figures and tables of Sickel et al. (in press) 186
Table 9 Individual author contributions for each part of Keller et al. (2016) 187
Table 10 Individual author contributions for the figures and tables of Keller et al. (2016) 187
Table 11 Individual author contributions for each part of Ankenbrand et al. (2017b) 188
Table 12 Individual author contributions for the figures and tables of Ankenbrand et al. (2017b) 188
Table 13 Individual author contributions for each part of Ankenbrand et al. (2017c) 189
Table 14 Individual author contributions for the figures and tables of Ankenbrand et al. (2017c) 189
Table 15 Individual author contributions for each part of Ankenbrand and Keller (2016) 190
Table 16 Individual author contributions for the figures and tables of Ankenbrand and Keller (2016) . . 190
Table 17 Individual author contributions for each part of the chloroExtractor manuscript 191
Table 18 Individual author contributions for the figures and tables of the chloroExtractor manuscript . 191
Table 19 Individual author contributions for each part of Ankenbrand et al. (2017a) 192
Table 20 Individual author contributions for the figures and tables of Ankenbrand et al. (2017a) 192

Table 21 Individual author contributions for each part of Ankenbrand et al. (2016) 193
Table 22 Individual author contributions for the figures and tables of Ankenbrand et al. (2016) 193

ACRONYMS

AliTV	Alignment Toolbox and Visualization
API	application programming interface
bcgTree	Bacterial Core Genome Tree
BIOM	Biological Observation Matrix
CBC	compensatory base change
DNA	deoxyribonucleic acid
FENNEC	Functional Exploration of Natural Networks and Ecological Communities
HMM	Hidden Markov Model
HTS	high-throughput sequencing
ITS2	Internal Transcribed Spacer 2
NCBI	The National Center for Biotechnology Information
OTU	operational taxonomic unit
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
rRNA	ribosomal RNA
SRA	Sequence Read Archive
TBro	Transcriptome Browser
TDD	Test Driven Development
tRNA	transfer RNA
USA	United States of America
WGA	whole genome alignment

Part I

INTRODUCTION

1 INCREASING VOLUME OF DATA IN BIOLOGY

1.1 THE BIOINFORMATIC BOTTLENECK

Recent technological advances led to the development of experimental methods with high throughput in biology (Greene et al., 2014). One of the most prominent high-throughput technologies is DNA sequencing with platforms like Illumina, PacBio and Oxford Nanopore (Metzker, 2010; Scholz et al., 2012). The availability of those methods opens a whole array of novel research areas (Dijk et al., 2014; Dolinski and Troyanskaya, 2015). Some say that there is now a "Fourth Paradigm" of data-intensive scientific discovery (the first three paradigms being empirical, theoretical and computational, Hey et al., 2009). The unprecedented pace in which new data is generated allows to tackle biological questions with novel approaches like DNA meta-barcoding, phylogenetic reconstruction from whole genomes, comparative genomics and transcriptome analysis of non-model organisms (Costa, 2014). However, along with the opportunities also many new challenges arise (Marx, 2013; Pop and Salzberg, 2008). Those include proper data storage, data preparation, visualization and statistical methods. Thus the bottleneck in current research projects is often shifted from cost of experiment to data analysis.

In order to analyze data from high-throughput experiments substantial bioinformatic knowledge and skills are required (Carvalho and Rustici, 2013). Bioinformatics is an integral part of biological and biomedical research beyond specialized analyses (Bork, 2005; Kanehisa and Bork, 2003). Some claim that computational biology is such an integral part of research that all biology is computational biology (Markowetz, 2017). An alternative to training all biologists in bioinformatics is making the data accessible through development of specialized software. This software can help cope with the amounts of data by providing interfaces to search, summarize and visualize the data. The slogan "Better software better research" coined by the Software Sustainability Institute (Crouch et al., 2013) illustrates that the quality of today's research depends on the quality of the used software (Goble, 2014).

1.2 CHALLENGES OF RESEARCH SOFTWARE ENGINEERING

Code has become an essential part of research. The amount of code ranges from a single script with few lines of code to whole frameworks with thousands of lines scattered over hundreds of files. In addition to the biological questions there are some technical and societal challenges to consider. Scientific software is often developed out of need by people not specifically trained in software development. Time constraints and the current scientific reward system lead to scripts that are hacked together quick and dirty. There has been a vivid discussion online whether scientific software should be viewed as a primary product of science at all (Brown, 2015a,b). So the code is often not easily reusable and error prone. Therefore code is not shared in many cases, leading to problems of reproducibility and multiple groups wasting efforts on the same (computational) problems (Rougier et al., 2017; Sandve et al., 2013). Furthermore errors in code can lead to wrong results and interpretations as exemplified by the retraction of multiple papers due to an error in a homemade data analysis script (Miller, 2006). Errors in more heavily used software can challenge or invalidate numerous studies. Such errors could be false assumptions in statistical packages for fMRI (Eklund et al., 2016) or automatic conversion of some gene names to dates in Microsoft Excel (Ziemann et al., 2016). It can be argued that there may be more undetected software errors that undermine scientific results (Soergel, 2015). Also many of the biological problems are computationally hard to solve and need specialized data structures and algorithms to be addressable (Ibsen-Jensen et al., 2015). Finding trained programmers to develop bioinformatic software is often also a problem as fundamental understanding of the biology is indispensable. In some cases scientists realize that the code they are writing for their project has broader applicability and decide to share it. Even in those cases the re-use is often limited as there is little time for proper testing and documentation of the software (Karimzadeh and Hoffman, 2017). Moreover it is in many cases a single person writing the code as part of a project or position (e.g. thesis) and as soon as this person moves on, the project is abandoned. Different computing environments further complicate broad application of existing software (Taschuk and Wilson, 2017). Finally, there is no standard repository for all scientific code. So finding the right tool for a bioinformatic task is not easy even if such a tool exists and is available. Those considerations have many practical implications on the design of the software projects of this thesis.

Although there has been substantial progress in bioinformatics, tool development is still lacking behind the enormous increase in biological data. The next chapter illustrates this fact by describing challenges, the current state, and limitations of applications in the fields of ecology, evolution, and genomics. The goal of this thesis is to develop and apply tools that help making more sense out of biological data in all those disciplines.

2 APPLICATIONS IN ECOLOGY, EVOLUTION, AND GENOMICS

The following sections contain examples of current bioinformatic applications in various biological fields. However, the scope is limited to topics relevant for this thesis, therefore there is no claim to completeness.

2.1 ECOLOGY

Ecology strives to unravel the factors that shape interactions between organisms among each other and with their environment. A mechanistic understanding of ecological patterns and processes is the goal of functional ecology (Irschick et al., 2013).

One of the most fundamental tasks of ecology is to determine the species composition of communities. Traditionally, this is done using morphological characteristics (Wiens and Servedio, 2000). high-throughput sequencing (HTS) facilitates usage of deoxyribonucleic acid (DNA) sequences to identify organisms instead (DNA barcoding). This method was first established for bacteria (Fox et al., 1977; Woese and Fox, 1977) but has been extended to eukaryots (Hebert et al., 2003). The genomic regions are called marker genes or barcodes and differ for taxonomic groups (Kress and Erickson, 2008). In DNA meta-barcoding the DNA is extracted from multiple organisms simultaneously to determine all of their members (Bálint et al., 2014; Yu et al., 2012). This novel approach solves a number of challenges of the traditional method (Hajibabaei et al., 2007). The most prominent advantages are more throughput, higher potential to detect rare species and deeper taxonomic assignments (Cowart et al., 2015). However, this procedure also has some limitations, especially regarding species abundance inference because of problems like primer bias and unequal biomass (Elbrecht and Leese, 2015). Also the huge amount of data produced by this method raises challenges (Coissac et al., 2012, Section 1.1). A typical workflow for meta-barcoding data processing consists of multiple steps. First the quality of the sequencing is checked with tools like FastQC (Andrews, 2010) and MultiQC (Ewels et al., 2016). Next the raw reads need to be processed (including demultiplexing, trimming, filtering and joining). For this process bioinformatic tools exist, e. g. QIIME (Caporaso et al., 2010), Trimmomatic (Bolger et al., 2014), usearch (Edgar, 2010), and fastq-join (Aronesty, 2013). The resulting sequences are compared against a reference database (e. g. BOLD (Ratnasingham and Hebert, 2007), RDP (Cole et al., 2005; Maidak et al., 1996)) with algorithms like

BLAST (Altschul et al., 1990), usearch (Edgar, 2010) or RDP classifier (Wang et al., 2007). Finally, the OTU tables or networks can be analyzed and visualized (e. g. vegan (Dixon, 2003), phyloseq (McMurdie and Holmes, 2013), Phinch (Bik and Pitch Interactive, 2014), Krona (Ondov et al., 2011)). There are even web tools that allow exploration of communities VAMPS (Huse et al., 2014) and MicrobiomeAnalyst (Dhariwal et al., 2017). Even a database with published community data exists in an early stage of development (Qiita (*Qiita* 2016)). So the full process of DNA meta-barcoding is well supported by bioinformatic tools. However, establishment of a meta-barcoding procedure in a new field of application requires adaptations to the standard protocol (Section 2.1.1). Furthermore, organismal properties (traits) are often missing from high throughput DNA meta-barcoding studies (Section 2.1.2).

2.1.1 *Pollen Meta-barcoding*

Pollination is an ecosystem service with economical relevance (Hanley et al., 2015). Therefore plant pollinator interactions are among the traditional study systems in ecology and evolution (Mitchell et al., 2009). One way to analyze this system is to inspect the pollen collected by individual bees. Traditionally, palynology involves time intensive identification of pollen under the light microscope (Mullins and Emberlin, 1997). This process requires expert knowledge and can often not happen to the species level due to morphological similarities of closely related taxa (Mullins and Emberlin, 1997). DNA meta-barcoding promises to overcome those problems (Galimberti et al., 2014). However, in order to adjust existing meta-barcoding procedures for pollen, multiple steps have to be taken. First selection of a marker gene, then establishment of an experimental protocol, finally the development of a bioinformatic workflow for data analysis. It is important to select a marker with suitable conservation and variability to be universally present but also to discriminate species (Hebert et al., 2003). Another important factor is the completeness and quality of the reference database. Popular marker genes for plants are Internal Transcribed Spacer 2 (ITS2), *rbcL*, *matK*, and *psbA-trnH* (Han et al., 2012; Kress et al., 2005). For the bioinformatic analysis a suitable reference database needs to be found and classifiers need to be trained. Furthermore an evaluation of the method compared to the traditional approach is required.

2.1.2 *Traits*

Comparing community structures on a taxonomic level can reveal interesting patterns. For example, that the effect of microhabitat filtering on plant-associated bacteria can exceed the influence of environmental effects (Junker and Keller, 2015). However, many biological questions can not be answered using taxonomy alone (Westoby and Wright, 2006). In some cases it is less important who is there but what they can do (Xu et al., 2014). A lot of morphological features and ecological functions have evolved in different taxonomic clades independently (Losos, 2011; Stern, 2013). In addition some features are secondarily lost and regained in some taxa, e.g. wings in insects (Stone and French, 2003; Whiting et al., 2003). So taxonomy does not correlate perfectly with traits (Junker et al., 2015). Thus it is not straightforward to answer ecologically relevant questions like what fractions of insects in the community can fly (e. g. as a measure for dispersal ability). In order to address this question the property "flying" has to be recorded for all organisms. This can be done either by direct observation, literature research, or by retrieving this value from a dedicated trait database. Other information about an organism like vulnerability or invasiveness are useful in biomonitoring and conservation biology (Keith et al., 2004). There is a fair amount of variation for some traits even inside one species (Forsman, 2015). Methods have been developed that use this phenotypic plasticity to analyze the functional composition of communities (Junker et al., 2016). This information can be used to describe and compare ecological niches (Winemiller et al., 2015). In summary DNA meta-barcoding can be used to address questions of functional trait evolution when combined with appropriate trait data (Kress et al., 2015; Uriarte et al., 2010).

Despite the fact that many public trait databases exist (e. g. TRY (Kattge et al., 2011), LEDA (Kleyer et al., 2008), IUCN (IUCN, 2017), BacDive (Söhngen et al., 2016), it is not easy to automatically retrieve trait data for whole communities. Common problems are that data formats are not standardized, information is scattered across multiple places, or there is no easy search or download for multiple organisms and traits. In addition the terms of data usage are often restrictive and do not allow free reuse. To solve some of those problems TraitBank (Parr et al., 2014b) aggregates traits from various sources and provides them via a unified web interface and an application programming interface (API). Similarly the rOpenSci traits module (Chamberlain et al., 2016) provides programmatic access to multiple trait sources from within R (R Core Team, 2017). Both tools provide some degree of standardization and machine readability but both are designed to retrieve trait information for individual species and not for whole communities. Further existing visualization tools like phyloSeq (McMurdie and Holmes, 2013) and Phinch (Bik and Pitch Interactive, 2014) are limited to handling taxonomic metadata for operational taxonomic units (OTUs). On the other hand the standard Biological Observation Matrix (BIOM) file format (McDonald et al., 2012) can store arbitrary OTU metadata. So a tool that can automate the trait aggregation and enrich existing communities would be useful. A simple user interface

(e. g. as a website) that handles BIOM data and allows for interactive exploration of trait patterns would facilitate usage by biologists.

2.2 EVOLUTION

Insights into evolutionary processes can be generated by reconstructing phylogenetic relationships. Knowledge of those relationships can help for example to track the spreading of a disease (Underwood et al., 2013) or to find treatments against pathogens (Hartfield et al., 2014). Furthermore, a proper phylogenetic analysis can be used to refine taxonomic assignments for ecological community analyses (Holt and Jønsson, 2014). Traditionally, trees are reconstructed using tables of morphological features (Hillis, 1987). Just like in classification DNA has emerged as a novel source of molecular features. In order to use DNA sequences for tree building, a genomic region (marker gene) has to be selected. A multiple sequence alignment of this region from each organism needs to be calculated in order to compare corresponding positions, with tools like CLUSTAL (Higgins and Sharp, 1988) or muscle (Edgar, 2004). Mature statistical methods for tree reconstruction (e. g. neighbor joining (Saitou and Nei, 1987), maximum parsimony (Fitch, 1971), maximum likelihood (Felsenstein, 1981), and Bayesian methods like Markov chain Monte Carlo (Yang and Rannala, 1997)) exist and are independent of the type of features (morphological or molecular). A whole array of software is available for the various methods, e. g. phylip (Felsenstein, 1989), RAxML (Stamatakis, 2014), and MrBayes (Huelsenbeck and Ronquist, 2001). The final trees can be visualized using tree view programs like figtree (Rambaut, 2017). Challenges are keeping specialized databases updated at the current pace of data generation (Section 2.2.1) and capitalizing on the increasing availability of whole genome data (Section 2.2.2).

2.2.1 *ITS2 Phylogeny*

The ITS2 is a commonly used marker for plants and fungi. It exhibits a variable sequence but a conserved secondary structure across eukaryots (Mai and Coleman, 1997; Schultz et al., 2005). Taking this secondary structure into account improves the accuracy and robustness of the resulting trees (Keller et al., 2010). Further a feature that can be detected in sequence structure alignments called compensatory base change (CBC), a mutation in two bases that maintains the bonding pattern, can be used to distinguish species (Müller et al., 2007). This observation holds despite intragenomic variation of the various copies of the ITS2 sequence inside one genome (Wolf et al., 2013). Specialized software for simultaneous sequence structure alignments (4SALE (Seibel et al., 2006, 2008), ProfDistS (Wolf et al., 2008), CBCanalyzer (Wolf et al., 2005a)) and a database with web based workbench for ITS2 sequences has been developed (Koetschan et al., 2010, 2012; Schultz et al., 2006). This database contains sequences that are crawled from The National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators, 2017), quality controlled, and re-annotated (Keller et al., 2009). With the ever increasing volume of sequences at NCBI keeping the ITS2 database up to date is an essential but non-trivial task.

2.2.2 *Multi-Marker Trees for Bacteria*

Beside the volume of sequences also the amount of full genomes, especially of bacteria in public databases is increasing (Kodama et al., 2012). As bacteria are hard to distinguish morphologically, DNA is the method of choice to classify them (Woese and Fox, 1977; Woese, 1987) Phylogenies can be reconstructed using a single or multiple markers (Ahrenfeldt et al., 2017; Queiroz and Gatesy, 2007). An appropriate genetic marker needs both sufficient information for distinguishing taxa and sufficient homology to make correct reconstructions (Capella-Gutierrez et al., 2014; Wu et al., 2013). One of the most commonly used markers is the 16S ribosomal RNA (rRNA) gene which has a high degree of conservation across prokaryots (Böttger, 1989; Clarridge, 2004). Often, different or multiple markers are used depending on the taxonomic level of the phylogenetic analysis to mitigate this trade-off between conservation and variability (Queiroz and Gatesy, 2007; Wu et al., 2013). Deriving a phylogeny from whole genomes is challenging, because there are often large genomic regions with no apparent similarities. Novichkov et al. (2009) address this challenge by providing a database of pre-calculated alignments of tight genomic clusters (ATGC). This enables high resolution micro-evolutionary analyses but is limited to the fraction of genomes included in their database. Alternatively instead of alignments of specific regions composition vectors of the whole genome can be used (Qi et al., 2004) as implemented in CVTree (Zuo and Hao, 2015). Another solution to the problem of marker selection is to concentrate on the conserved regions present in a majority of organisms of interest (Ciccarelli et al., 2006). But even with a defined set of target genes it requires substantial bioinformatic skills to find, extract, and align them from whole bacterial genomes. An automatic procedure for this extraction, as well as combination of the multi marker alignments and tree reconstruction would be useful.

2.3 GENOMICS

Genomes are the blueprints of organisms. Thus expectations in genome projects are generally high (Iliopoulos et al., 2001). Deciphering the genome sequence of an organism can help in understanding how they are performing their unique functions, how they are adapted to their ecological niche and possibly also how evolution shaped it. Through

HTS technologies genome projects became affordable (Pareek et al., 2011). Nowadays also large and complex eukaryotic genomes (like plants) are routinely analyzed. However, completion of a genome sequence is only the first step in really understanding its structure and function (Butler and Smaglik, 2000). A current genome project involves among others quality control, read error correction, genome assembly, and gene annotation. Bioinformatic tools have been developed to support all of those steps. Quality of sequences can be assessed with FastQC (Andrews, 2010) or MultiQC (Ewels et al., 2016). There are different approaches to error correction for short reads (e. g. ECHO (Kao et al., 2011)) and for long reads (e. g. proovread (Hackl et al., 2014)). Assemblers facilitating a set of methods (e. g. overlap/layout/consensus and de Bruijn graphs) have been developed (Miller et al., 2010). One of the most commonly uses assemblers for bacteria is SPAdes which uses a de Bruijn graph approach and integrates an error correction step (Bankevich et al., 2012) Dedicated annotation tools exist for functional protein annotation (e. g. MAKER (Cantarel et al., 2008)), rRNA genes (e. g. RNAmmer (Lagesen et al., 2007)), transfer RNA (tRNA) genes (e. g. ARAGORN (Laslett and Canback, 2004)), and repeats (e. g. RepeatMasker (Smit et al., 2013)). With many complete genomes being published in recent years (O'Leary et al., 2016) it becomes feasible to compare differences in genome architecture on the large scale. Also fueled by HTS the analysis of transcriptomes via RNA sequencing (RNA-Seq) was established (Nagalakshmi et al., 2008; Wang et al., 2009). This way also non-model organisms for which no reference genome is available can be examined. The next sections describe the challenges of extracting plastid genomes from whole genome sequencing reads (Section 2.3.1), visualizing whole genome alignments (Section 2.3.2), and exploring *de novo* transcriptome data (Section 2.3.3).

2.3.1 *Plastid Genomes*

Beside the nuclear genomes plants also have plastid and mitochondrial genomes. Chloroplast genomes are interesting in their own right (Daniell et al., 2016). In addition whole chloroplasts can be used for phylogenetic reconstruction (Huang et al., 2016) and even barcoding (Coissac et al., 2016). Specialized tools for the annotation of chloroplasts have been developed (e. g. DOGMA (Wyman et al., 2004) and CpGAVAS (Liu et al., 2012a)). Those annotations are commonly visualized as circular maps using OGDRAW (Lohse et al., 2013). Resources for storage and comparison of genome sequences and annotation exist (e. g. ChloroplastDB (Cui et al., 2006) and Verdant (McKain et al., 2017)). Nowadays, chloroplast genomes are often not specifically sequenced but appear as a byproduct of whole genome sequencing. In order to analyze the genome and the chloroplast individually their sequences need to be separated. However, this requires sophisticated bioinformatic methods. Therefore, to benefit from the chloroplast sequences, e. g. for comparative genomics, a tool that automatically extracts and assembles the chloroplast from raw data sets would be useful.

2.3.2 *Comparative Genomics*

The publicly available genomes can be explored for large-scale evolutionary processes using comparative genomics. They can be used to study genomic recombination (Didelot et al., 2012), horizontal gene transfer and genomic islands (Avrani et al., 2011; Langille et al., 2008) as well as the dispersal of viral elements (Touchon and Rocha, 2007). To find large scale differences in genome architecture, whole genome alignments (WGAs) are used (Couronne et al., 2003). Those can be calculated with tools like lastz (*LASTZ* 2015) or MUMmer 2 (Delcher et al., 2002). A major challenge is the interpretation of those WGAs. Without visualization of the results the patterns are hard to grasp. So a collection of tools providing graphical representations of aligned genomes have been developed over the years (e. g. Mauve (Darling et al., 2004), ACT (Carver et al., 2008), genoPlotR (Guy et al., 2010), BRIG (Alikhan et al., 2011), and EasyFig (Sullivan et al., 2011)). Those programs together cover a broad range of user interfaces and visualizations. However, no single one of them combines a linear representation (allowing for multiple alignments), interactive modification of parameters (e. g. filtering, rotation, arrangement), and high quality export of the final graphic. Particularly the interactive exploration helps to spot hidden patterns and allows for more detailed analyses.

2.3.3 *De-Novo Transcriptomics*

In contrast to the genome the transcriptome contains only genes that are active in a given tissue at a given time (Wang et al., 2009). Understanding the transcriptome is essential for deciphering the functional properties of the genome and how it shapes different cell types and tissues. It also aids in understanding developmental processes and diseases (Wang et al., 2007). Transcriptome analysis differs depending on whether or not a reference genome is available. In the first case sequencing reads are usually mapped on the reference genome (e. g. with bowtie2 (Langmead and Salzberg, 2012)) and existing annotations are used for quantification. In the latter case the reads are first assembled (e. g. with Trinity (Grabherr et al., 2011) or Oases (Schulz et al., 2012)). The resulting transcripts are quantified with tools like RSEM (Li and Dewey, 2011) or sailfish (Patro et al., 2014) and annotated with programs like interproScan (Jones et al., 2014), Mercator (Lohse et al., 2014), and blast2go (Conesa and Götz, 2008). If counts are available for multiple tissues or conditions, differential gene expression can be statistically tested with tools like DESeq2 (Love et al., 2014) or edgeR

(Robinson et al., 2010). For eukaryots transcript counts can go into the hundred thousands and thus tools for exploratory analysis, subsetting and visualization are essential. Genome browsers can be utilized in case of reference based transcriptomics. However, no equivalent option exists for *de novo* transcriptomics.

3 AIM OF THIS THESIS

The aim of this thesis is to facilitate biological research by developing and applying bioinformatic tools. In the previous chapter multiple challenges in ecology, evolution and genomics have been identified. Despite the fact that a lot of algorithms and programs exist to solve parts of those challenges, there is still room for improvement. Moreover the different fields of biology are not isolated. For example, taking aspects of evolution and genomics into account gives a more holistic picture of ecosystem function. Thus the tools developed in different areas aim for integrative analyses combining data from different fields. Another main goal is the utilization of publicly available data for novel applications. The software should facilitate re-use and further development by adhering to best practices. Through good usability the software should enable biologists to answer research questions with massive amounts of data.

Succinctly the aim of this thesis is to facilitate research that would not be possible without it.

Part II

MATERIAL AND METHODS

4 SOURCE CODE

The materials and methods used in this thesis are described in detail in each of the individual publications in Chapter 5. However, there have been some general methods used throughout all of the projects to ensure code quality and the reproducibility of analyses. In general all code written for this thesis is openly available at GitHub (`https://github.com/molbiodiv`). This includes scripts and instructions to re-run analyses as well as source code of bioinformatic tools. All the material is licensed permissively under the terms of the MIT License. GitHub facilitates usage and collaboration by providing standard mechanisms to report issues, fork code and contribute via pull requests. All code for software projects is developed under the Test Driven Development (TDD) paradigm. This means that new features are added in cycles of writing tests, seeing them fail, implement the new feature until tests pass, and then refactoring the code (for better readability and/or performance). Continuous integration via Travis CI ensures that all tests pass in a defined and clean environment. All tools are documented with at least simple usage examples. Some even more extensively with comprehensive tutorials and demo data. For server applications ready to use docker images are provided via DockerHub. One important thing to note is that many of the software projects have been collaborative efforts so not every line of code has been written by me. Usually the author lists of the corresponding papers provide fairly accurate representations of individual contributions. For more details the git repositories can be explored. Here I provide a brief summary of each repository associated with this thesis. The list of programming languages, libraries, and frameworks for a project is limited on the most important ones for the project and might not be extensive.

4.1 META-BARCODING-DUAL-INDEXING

This project contains the reproducible workflow for the pollen metabarcoding analysis in Sickel et al. (2015). It also contains a detailed description of training the classifiers with own data. All required data as well as pre-computed results are provided. The material covers large parts of the bioinformatic methodology of Keller et al. (2015).

REPOSITORY
`https://github.com/molbiodiv/meta-barcoding-dual-indexing`

PROGRAMMING LANGUAGES
Perl, Bash

EXTERNAL PROGRAMS
RDP Classifier (Wang et al., 2007), usearch (Edgar, 2010), QIIME (Caporaso et al., 2010)

SOURCE CODE DOI
10.5281/zenodo.61726

4.2 BIOJS-IO-BIOM

A generic JavaScript library to handle data in BIOM format (McDonald et al., 2012). This is used to store user projects in Functional Exploration of Natural Networks and Ecological Communities (FENNEC), it is described in Ankenbrand et al. (2017b).

REPOSITORY
https://github.com/molbiodiv/biojs-io-biom

PROGRAMMING LANGUAGES
JavaScript (ES6)

LIBRARIES AND FRAMEWORKS
lodash (https://lodash.com/), BioJS (Corpas et al., 2014), mocha (https://mochajs.org/)

CONTINUOUS INTEGRATION
https://travis-ci.org/molbiodiv/biojs-io-biom

SOURCE CODE DOI
10.5281/zenodo.597920

4.3 BIOM-CONVERSION-SERVER

A lightweight server interface to the biom-format command line tool (McDonald et al., 2012). It provides conversion capability between BIOM version 1 and 2. The biojs-io-biom library (Ankenbrand et al., 2017b) uses this server to handle the binary HDF5 file format.

REPOSITORY
https://github.com/molbiodiv/biom-conversion-server

PROGRAMMING LANGUAGES
PHP

LIBRARIES AND FRAMEWORKS

biom-format (python) (McDonald et al., 2012), biojs-io-biom (Ankenbrand et al., 2017b)

CONTINUOUS INTEGRATION

https://travis-ci.org/molbiodiv/biom-conversion-server

DOCKER

https://hub.docker.com/r/iimog/biom-conversion-server/

SOURCE CODE DOI

10.5281/zenodo.597903

4.4 FENNEC

The code base for the FENNEC web application. This software stores traits aggregated from various databases and automatically enriches user provided community tables (Ankenbrand et al., 2017c). It also includes a modified version of Phinch (Bik and Pitch Interactive, 2014) for interactive visualization of trait distributions.

REPOSITORY

https://github.com/molbiodiv/fennec

PROGRAMMING LANGUAGES

PHP, HTML, CSS, JavaScript

LIBRARIES AND FRAMEWORKS

Symfony (https://symfony.com/), doctrine (http://doctrine-project.org/), HWIOAuthBundle (https://github.com/hwi/HWIOAuthBundle), react (https://facebook.github.io/react/), plotly.js (https://plot.ly/javascript/), jQuery (https://jquery.com/), lodash (https://lodash.com/), phpunit (https://phpunit.de/), mocha (https://mochajs.org/), webpack (https://webpack.github.io/), babel (https://babeljs.io/)

DOCUMENTATION

http://fennec.readthedocs.io/en/latest/

CONTINUOUS INTEGRATION

https://travis-ci.org/molbiodiv/fennec

DOCKER

https://hub.docker.com/r/iimog/fennec/

PUBLIC INSTANCE
http://fennec.molecular.eco

SOURCE CODE DOI
10.5281/zenodo.591305

4.5 BCGTREE

The source code for both the command line tool and the graphical user interface to reconstruct phylogenetic trees from bacterial core genomes (Ankenbrand and Keller, 2016). It also includes the Hidden Markov Model (HMM) models for the 107 essential genes (Dupont et al., 2012).

REPOSITORY
https://github.com/molbiodiv/bcgTree

PROGRAMMING LANGUAGES
Perl, Java

EXTERNAL PROGRAMS
HMMER (Eddy, 2010), muscle (Edgar, 2004), Gblocks (Castresana, 2000)

CONTINUOUS INTEGRATION
https://travis-ci.org/molbiodiv/bcgTree

SOURCE CODE DOI
10.5281/zenodo.597913

4.6 CHLOROEXTRACTOR

Source code of the chloroExtractor allowing for automatic extraction and assembly of chloroplast genomes from whole genome shotgut data.

REPOSITORY
https://github.com/chloroExtractorTeam/chloroExtractor

PROGRAMMING LANGUAGES
Perl

LIBRARIES AND FRAMEWORKS
jellyfish (Marçais and Kingsford, 2011), bowtie 2 (Langmead and Salzberg, 2012), Spades (Bankevich et al., 2012)

CONTINUOUS INTEGRATION
`https://travis-ci.org/chloroExtractorTeam/chloroExtractor`

SOURCE CODE DOI
10.5281/zenodo.883594

4.7 ALITV

The source code for the Alignment Toolbox and Visualization (AliTV) is split into a perl part to generate json files from multiple alignments and a JavaScript part that visualizes the results interactively in a web-browser. AliTV is described in detail in Ankenbrand et al. (2017a).

REPOSITORIES
`https://github.com/AliTVTeam/AliTV`,
`https://github.com/AliTVTeam/AliTV-perl-interface`

PROGRAMMING LANGUAGES
Perl, JavaScript

LIBRARIES AND FRAMEWORKS
jasmine (`https://jasmine.github.io/`), d3 (`https://d3js.org/`), jQuery (`https://jquery.com/`), underscore (`http://underscorejs.org/`)

EXTERNAL PROGRAMS
lastz (Harris, 2007)

DOCUMENTATION
`http://alitv.readthedocs.io/en/latest/`

CONTINUOUS INTEGRATION
`https://travis-ci.org/AliTVTeam/`

PUBLIC INSTANCE
`https://alitvteam.github.io/AliTV/d3/AliTV.html`

SOURCE CODE DOIS
10.5281/zenodo.597917, 10.5281/zenodo.597916

4.8 TBRO

Source code of Transcriptome Browser (TBro) is split across multiple repositories. It contains the server part, docker containers, demo data, documentation and the command line interface in separate repositories. The TBro is a web based workbench for the analysis of *de novo* transcriptome studies (Ankenbrand et al., 2016).

REPOSITORIES

https://github.com/TBroTeam/

PROGRAMMING LANGUAGES

PHP, HTML, CSS, JavaScript

LIBRARIES AND FRAMEWORKS

Smarty (https://www.smarty.net/), Foundation (https://foundation.zurb.com/), underscore (http://underscorejs.org/), jQuery (https://jquery.com/), canvasXpress (Neuhaus, 2016)

DOCUMENTATION

http://tbro-tutorial.readthedocs.io/en/latest/

CONTINUOUS INTEGRATION

https://travis-ci.org/TBroTeam/TBro

DOCKER

https://hub.docker.com/u/tbroteam/

SOURCE CODE DOI

10.5281/zenodo.597901

Part III

RESULTS

5 PUBLICATIONS

This chapter includes the publications listed on page vii. Original publisher PDF pages or supplementary material are indicated by frames enclosing the respective pages. In two cases with open peer review the reviewer reports and our responses are included as well.

5.1 ITS2 DATABASE V: TWICE AS MUCH

– published in *Molecular Biology and Evolution* –

ITS2 Database V: Twice as Much

Markus J. Ankenbrand,[1] Alexander Keller,[1] Matthias Wolf,[2] Jörg Schultz,[2] and Frank Förster*[,2]

[1]Department of Animal Ecology and Tropical Biology, Julius Maximilian University, Würzburg, Germany

[2]Department of Bioinformatics, Julius Maximilian University, Würzburg, Germany

***Corresponding author:** E-mail: frank.foerster@biozentrum.uni-wuerzburg.de.

Associate editor: Koichiro Tamura

Abstract

The internal transcribed spacer 2 (ITS2) is a well-established marker for phylogenetic analyses in eukaryotes. A reliable resource for reference sequences and their secondary structures is the ITS2 database (http://its2.bioapps.biozentrum.uni-wuerzburg.de/). However, the database was last updated in 2011. Here, we present a major update of the underlying data almost doubling the number of entities. This increases the number of taxa represented within all major eukaryotic clades. Moreover, additional data has been added to underrepresented groups and some new groups have been added. The broader coverage across the tree of life improves phylogenetic analyses and the capability of ITS2 as a DNA barcode.

Key words: **barcoding, database, internal transcribed spacer, phylogeny, sequence-structure, toolbox.**

Introduction

The internal transcribed spacer 2 (ITS2) of the ribosomal cistron is a well-established marker in eukaryotic molecular systematics (Schultz and Wolf 2009). With a relatively variable sequence it is well suited for low-level analyses, yet limited for distantly related taxa (Baldwin 1992). However, ITS2 exhibits a common core of secondary structure (Schultz et al. 2005) making it a valuable marker also on higher taxonomic levels (Coleman 2003). Furthermore, inclusion of the secondary structure improves the accuracy and robustness of phylogenetic tree reconstructions (Keller et al. 2010) and allows for distinguishing cryptic/pseudocryptic species via compensatory base changes (Müller et al. 2007; Coleman 2009; Ruhl et al. 2010). Recently, it has also been applied in DNA (meta-) barcoding (Chen et al. 2010; Yao et al. 2010; Pang et al. 2012; Keller et al. 2015).

In 2006, we developed the ITS2 database to provide a central resource for ITS2 sequences and their individual secondary structures (Schultz et al. 2006). In the following years, the ITS2 database was further expanded from a data repository to a rather full featured interactive workbench (Selig et al. 2008; Koetschan et al. 2010, 2012; Wolf et al. 2014). Data of the ITS2 workbench consist of sequences extracted from NCBI (NCBI Resource Coordinators 2015) that are automatically trimmed using Hidden Markov Models (Keller et al. 2009). The workbench determines complete individual secondary structures for ITS2 sequences based on energy minimization (Markham and Zuker 2008) or iterative homology modelling (Wolf et al. 2005). Additionally, partial structures are predicted for entries with as few as two helices (Koetschan et al. 2010). Finally, ITS2 sequences without a predicted structure are included as sequence-only entities (Koetschan et al. 2010). During the automatic structure validation all entries have to match the four helix core. Thus, other ITS2 structures are not represented in our database. Basic analyses like reannotation, secondary structure prediction, sequence-structure alignment, and tree calculation can be directly performed in the web-based database (Merget et al. 2012). The last update of the underlying data was performed in 2011. Meanwhile, the NCBI database experienced a drastic increase in sequence content (supplementary table S1, Supplementary Material online). Moreover, the NCBI Taxonomy (Federhen 2012) is continuously revised to reflect the current knowledge of the evolutionary history of represented taxa. We thus performed a major update on the ITS2 workbench to benefit from this increased amount of data and make it available to the scientific ITS2 communities.

In the following, we report the most prominent improvements in terms of stored data, taxonomic coverage, and changes in major lineages.

Results

The new version of the database now contains 711,172 sequences, which nearly doubles the 379,329 of the previous release. In detail, the number of entries matching the eukaryotic core structure increased by 84%, and those with a partial structure increased by 217%. In contrast, the number of sequences without structure decreased by 11%. Similarly, the number of different species and genera represented in the database increased by 59% and 23%, respectively. Overall the proportional increase in number of new sequences was distributed across all major groups of eukaryotes (table 1).

The taxonomic lineage for each sequence was updated to the current NCBI Taxonomy and also showed some major changes. The NCBI TaxIDs for 7,464 sequences were changed since the last update. 3,743 entries present in 2011 are altered in the current update (supplementary table S2, Supplementary Material online).

Discussion

When calculating reliable phylogenetic trees or when performing DNA barcoding analyses, it is essential to have a

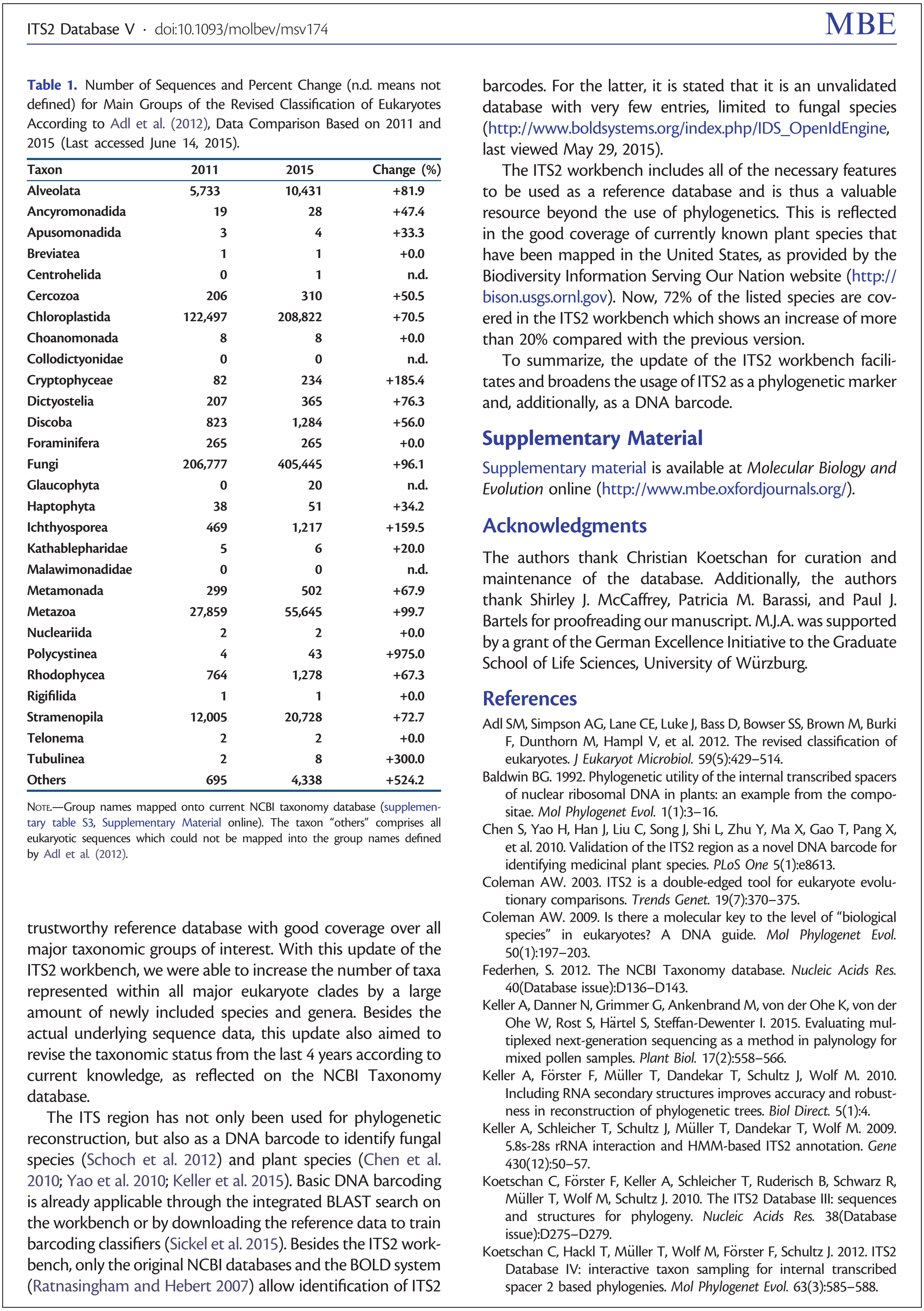


Table 1. Number of Sequences and Percent Change (n.d. means not defined) for Main Groups of the Revised Classification of Eukaryotes According to Adl et al. (2012), Data Comparison Based on 2011 and 2015 (Last accessed June 14, 2015).

Taxon	2011	2015	Change (%)
Alveolata	5,733	10,431	+81.9
Ancyromonadida	19	28	+47.4
Apusomonadida	3	4	+33.3
Breviatea	1	1	+0.0
Centrohelida	0	1	n.d.
Cercozoa	206	310	+50.5
Chloroplastida	122,497	208,822	+70.5
Choanomonada	8	8	+0.0
Collodictyonidae	0	0	n.d.
Cryptophyceae	82	234	+185.4
Dictyostelia	207	365	+76.3
Discoba	823	1,284	+56.0
Foraminifera	265	265	+0.0
Fungi	206,777	405,445	+96.1
Glaucophyta	0	20	n.d.
Haptophyta	38	51	+34.2
Ichthyosporea	469	1,217	+159.5
Kathablepharidae	5	6	+20.0
Malawimonadidae	0	0	n.d.
Metamonada	299	502	+67.9
Metazoa	27,859	55,645	+99.7
Nucleariida	2	2	+0.0
Polycystinea	4	43	+975.0
Rhodophycea	764	1,278	+67.3
Rigifilida	1	1	+0.0
Stramenopila	12,005	20,728	+72.7
Telonema	2	2	+0.0
Tubulinea	2	8	+300.0
Others	695	4,338	+524.2

NOTE.—Group names mapped onto current NCBI taxonomy database (supplementary table S3, Supplementary Material online). The taxon "others" comprises all eukaryotic sequences which could not be mapped into the group names defined by Adl et al. (2012).

trustworthy reference database with good coverage over all major taxonomic groups of interest. With this update of the ITS2 workbench, we were able to increase the number of taxa represented within all major eukaryote clades by a large amount of newly included species and genera. Besides the actual underlying sequence data, this update also aimed to revise the taxonomic status from the last 4 years according to current knowledge, as reflected on the NCBI Taxonomy database.

The ITS region has not only been used for phylogenetic reconstruction, but also as a DNA barcode to identify fungal species (Schoch et al. 2012) and plant species (Chen et al. 2010; Yao et al. 2010; Keller et al. 2015). Basic DNA barcoding is already applicable through the integrated BLAST search on the workbench or by downloading the reference data to train barcoding classifiers (Sickel et al. 2015). Besides the ITS2 workbench, only the original NCBI databases and the BOLD system (Ratnasingham and Hebert 2007) allow identification of ITS2 barcodes. For the latter, it is stated that it is an unvalidated database with very few entries, limited to fungal species (http://www.boldsystems.org/index.php/IDS_OpenIdEngine, last viewed May 29, 2015).

The ITS2 workbench includes all of the necessary features to be used as a reference database and is thus a valuable resource beyond the use of phylogenetics. This is reflected in the good coverage of currently known plant species that have been mapped in the United States, as provided by the Biodiversity Information Serving Our Nation website (http://bison.usgs.ornl.gov). Now, 72% of the listed species are covered in the ITS2 workbench which shows an increase of more than 20% compared with the previous version.

To summarize, the update of the ITS2 workbench facilitates and broadens the usage of ITS2 as a phylogenetic marker and, additionally, as a DNA barcode.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Christian Koetschan for curation and maintenance of the database. Additionally, the authors thank Shirley J. McCaffrey, Patricia M. Barassi, and Paul J. Bartels for proofreading our manuscript. M.J.A. was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.

References

Adl SM, Simpson AG, Lane CE, Luke J, Bass D, Bowser SS, Brown M, Burki F, Dunthorn M, Hampl V, et al. 2012. The revised classification of eukaryotes. *J Eukaryot Microbiol.* 59(5):429–514.

Baldwin BG. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the compositae. *Mol Phylogenet Evol.* 1(1):3–16.

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, et al. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5(1):e8613.

Coleman AW. 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.* 19(7):370–375.

Coleman AW. 2009. Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. *Mol Phylogenet Evol.* 50(1):197–203.

Federhen, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res.* 40(Database issue):D136–D143.

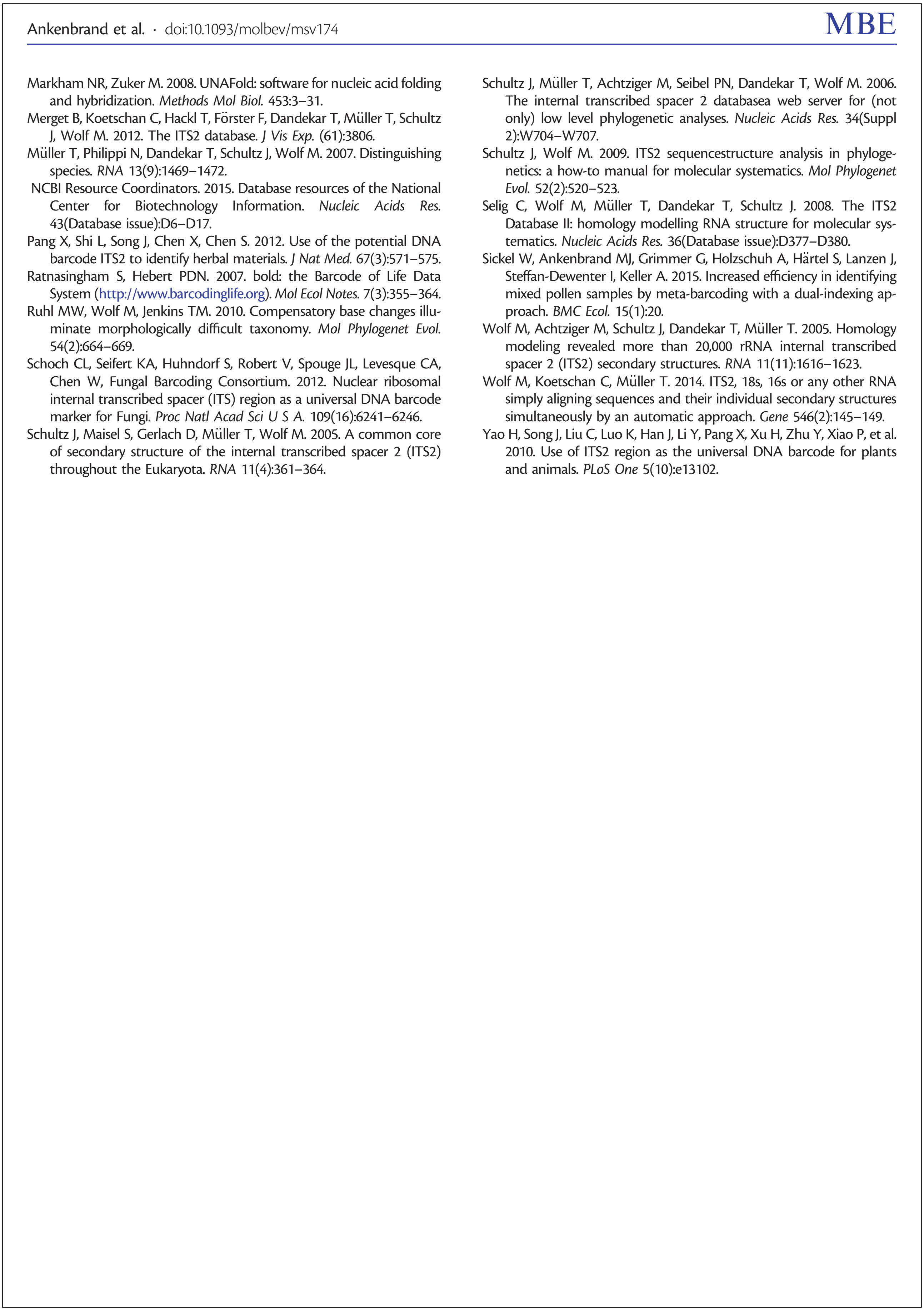
Keller A, Danner N, Grimmer G, Ankenbrand M, von der Ohe K, von der Ohe W, Rost S, Härtel S, Steffan-Dewenter I. 2015. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biol.* 17(2):558–566.

Keller A, Förster F, Müller T, Dandekar T, Schultz J, Wolf M. 2010. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biol Direct.* 5(1):4.

Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M. 2009. 5.8s-28s rRNA interaction and HMM-based ITS2 annotation. *Gene* 430(12):50–57.

Koetschan C, Förster F, Keller A, Schleicher T, Ruderisch B, Schwarz R, Müller T, Wolf M, Schultz J. 2010. The ITS2 Database III: sequences and structures for phylogeny. *Nucleic Acids Res.* 38(Database issue):D275–D279.

Koetschan C, Hackl T, Müller T, Wolf M, Förster F, Schultz J. 2012. ITS2 Database IV: interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Mol Phylogenet Evol.* 63(3):585–588.



Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol.* 453:3–31.

Merget B, Koetschan C, Hackl T, Förster F, Dandekar T, Müller T, Schultz J, Wolf M. 2012. The ITS2 database. *J Vis Exp.* (61):3806.

Müller T, Philippi N, Dandekar T, Schultz J, Wolf M. 2007. Distinguishing species. *RNA* 13(9):1469–1472.

NCBI Resource Coordinators. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 43(Database issue):D6–D17.

Pang X, Shi L, Song J, Chen X, Chen S. 2012. Use of the potential DNA barcode ITS2 to identify herbal materials. *J Nat Med.* 67(3):571–575.

Ratnasingham S, Hebert PDN. 2007. bold: the Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes.* 7(3):355–364.

Ruhl MW, Wolf M, Jenkins TM. 2010. Compensatory base changes illuminate morphologically difficult taxonomy. *Mol Phylogenet Evol.* 54(2):664–669.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A.* 109(16):6241–6246.

Schultz J, Maisel S, Gerlach D, Müller T, Wolf M. 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA* 11(4):361–364.

Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M. 2006. The internal transcribed spacer 2 databasea web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.* 34(Suppl 2):W704–W707.

Schultz J, Wolf M. 2009. ITS2 sequencestructure analysis in phylogenetics: a how-to manual for molecular systematics. *Mol Phylogenet Evol.* 52(2):520–523.

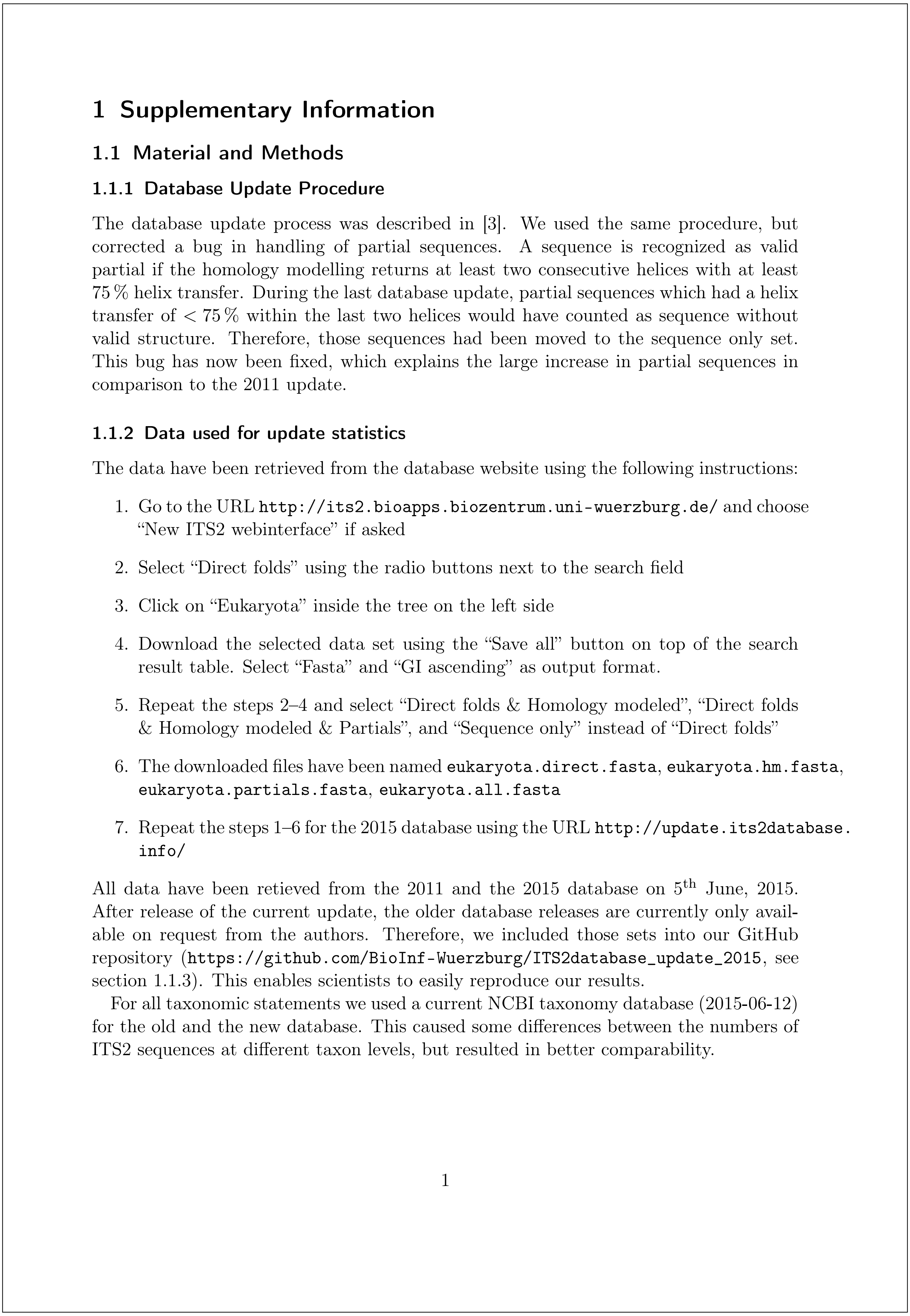
Selig C, Wolf M, Müller T, Dandekar T, Schultz J. 2008. The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Res.* 36(Database issue):D377–D380.

Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, Steffan-Dewenter I, Keller A. 2015. Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.* 15(1):20.

Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T. 2005. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA* 11(11):1616–1623.

Wolf M, Koetschan C, Müller T. 2014. ITS2, 18s, 16s or any other RNA simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene* 546(2):145–149.

Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, et al. 2010. Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* 5(10):e13102.



1 Supplementary Information

1.1 Material and Methods

1.1.1 Database Update Procedure

The database update process was described in [3]. We used the same procedure, but corrected a bug in handling of partial sequences. A sequence is recognized as valid partial if the homology modelling returns at least two consecutive helices with at least 75 % helix transfer. During the last database update, partial sequences which had a helix transfer of $< 75\,\%$ within the last two helices would have counted as sequence without valid structure. Therefore, those sequences had been moved to the sequence only set. This bug has now been fixed, which explains the large increase in partial sequences in comparison to the 2011 update.

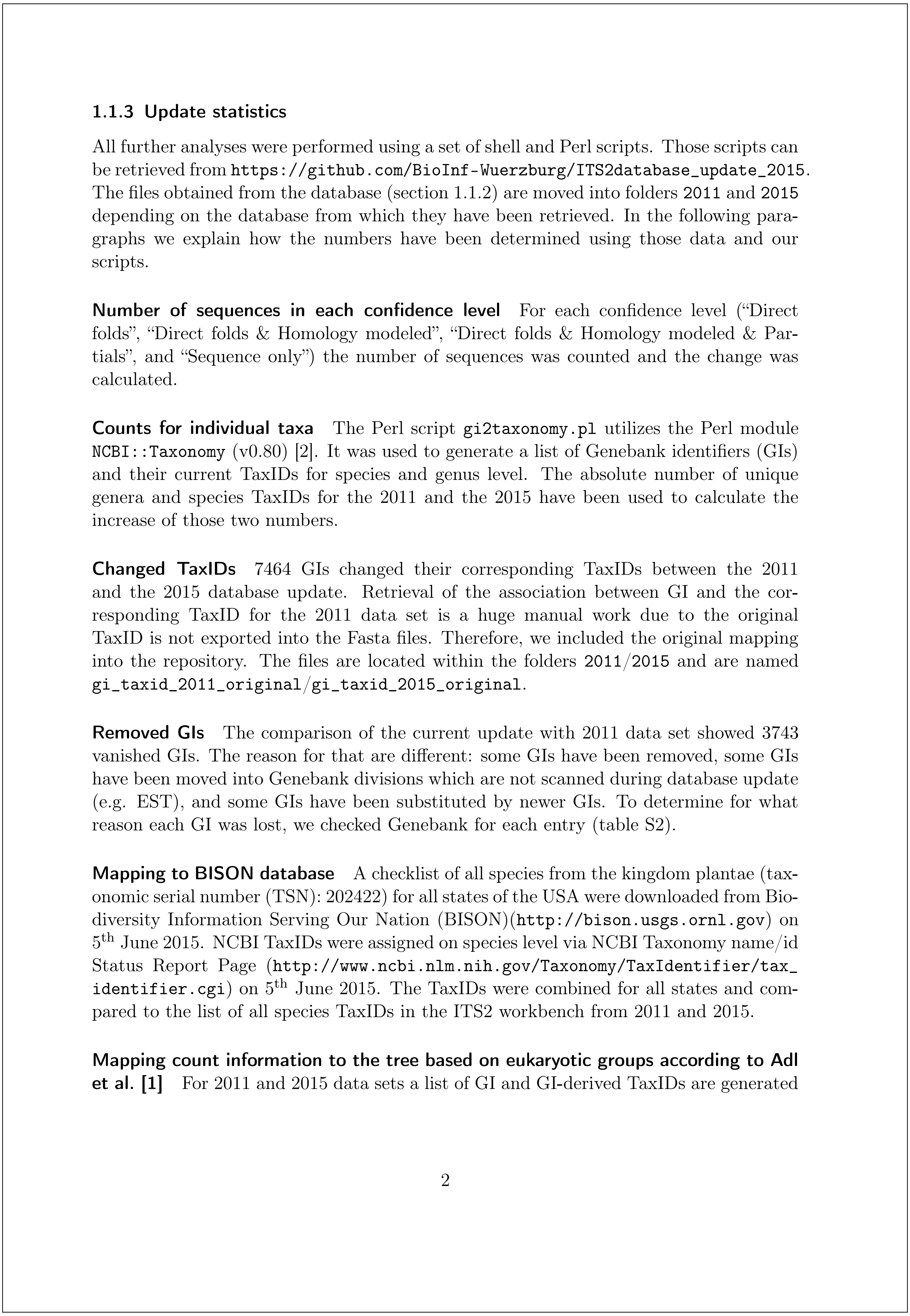
1.1.2 Data used for update statistics

The data have been retrieved from the database website using the following instructions:

1. Go to the URL `http://its2.bioapps.biozentrum.uni-wuerzburg.de/` and choose “New ITS2 webinterface” if asked
2. Select “Direct folds” using the radio buttons next to the search field
3. Click on “Eukaryota” inside the tree on the left side
4. Download the selected data set using the “Save all” button on top of the search result table. Select “Fasta” and “GI ascending” as output format.
5. Repeat the steps 2–4 and select “Direct folds & Homology modeled”, “Direct folds & Homology modeled & Partials”, and “Sequence only” instead of “Direct folds”
6. The downloaded files have been named `eukaryota.direct.fasta`, `eukaryota.hm.fasta`, `eukaryota.partials.fasta`, `eukaryota.all.fasta`
7. Repeat the steps 1–6 for the 2015 database using the URL `http://update.its2database.info/`

All data have been retieved from the 2011 and the 2015 database on 5th June, 2015. After release of the current update, the older database releases are currently only available on request from the authors. Therefore, we included those sets into our GitHub repository (`https://github.com/BioInf-Wuerzburg/ITS2database_update_2015`, see section 1.1.3). This enables scientists to easily reproduce our results.

For all taxonomic statements we used a current NCBI taxonomy database (2015-06-12) for the old and the new database. This caused some differences between the numbers of ITS2 sequences at different taxon levels, but resulted in better comparability.



1.1.3 Update statistics

All further analyses were performed using a set of shell and Perl scripts. Those scripts can be retrieved from `https://github.com/BioInf-Wuerzburg/ITS2database_update_2015`. The files obtained from the database (section 1.1.2) are moved into folders `2011` and `2015` depending on the database from which they have been retrieved. In the following paragraphs we explain how the numbers have been determined using those data and our scripts.

Number of sequences in each confidence level For each confidence level ("Direct folds", "Direct folds & Homology modeled", "Direct folds & Homology modeled & Partials", and "Sequence only") the number of sequences was counted and the change was calculated.

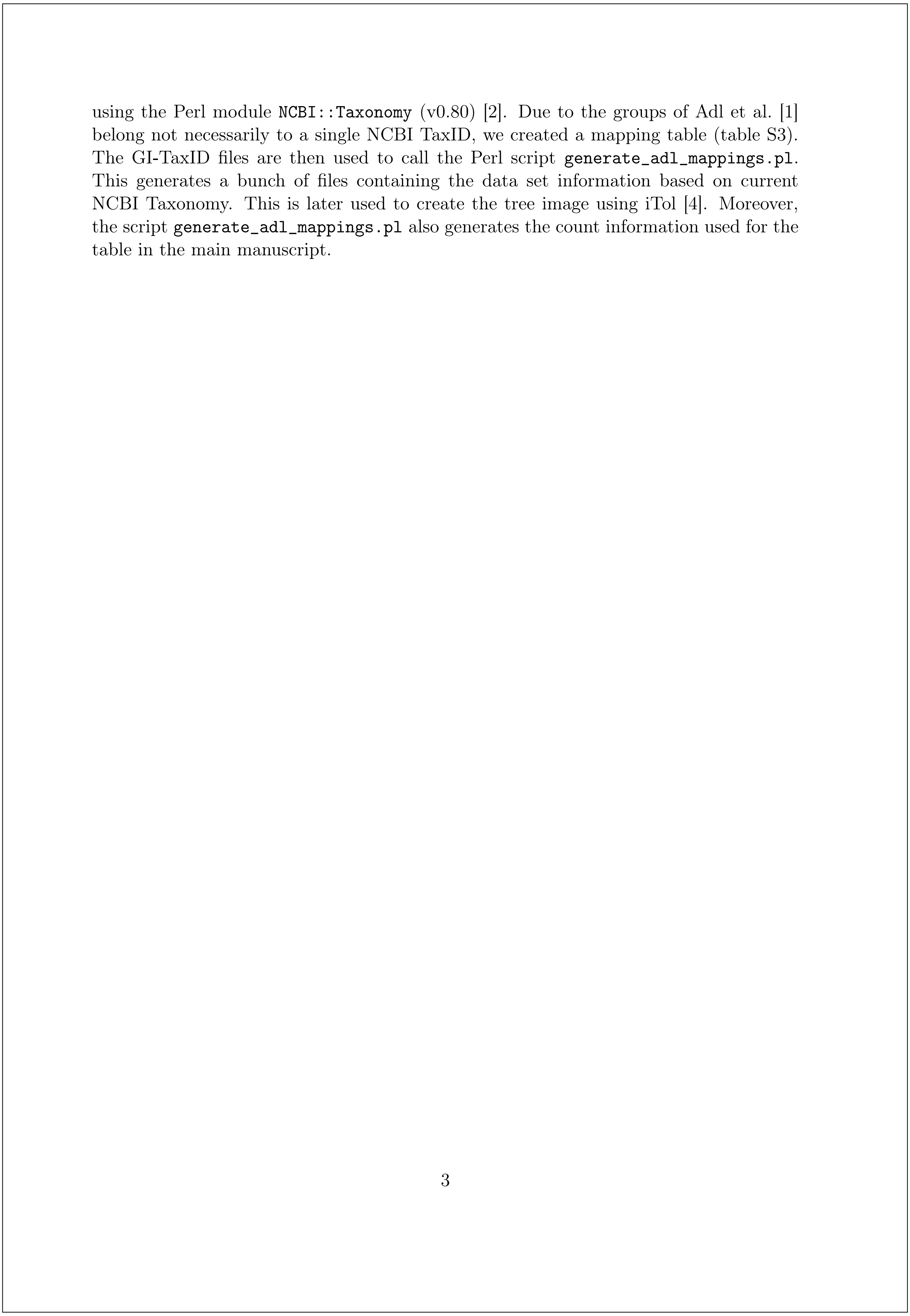
Counts for individual taxa The Perl script `gi2taxonomy.pl` utilizes the Perl module `NCBI::Taxonomy` (v0.80) [2]. It was used to generate a list of Genebank identifiers (GIs) and their current TaxIDs for species and genus level. The absolute number of unique genera and species TaxIDs for the 2011 and the 2015 have been used to calculate the increase of those two numbers.

Changed TaxIDs 7464 GIs changed their corresponding TaxIDs between the 2011 and the 2015 database update. Retrieval of the association between GI and the corresponding TaxID for the 2011 data set is a huge manual work due to the original TaxID is not exported into the Fasta files. Therefore, we included the original mapping into the repository. The files are located within the folders `2011`/`2015` and are named `gi_taxid_2011_original`/`gi_taxid_2015_original`.

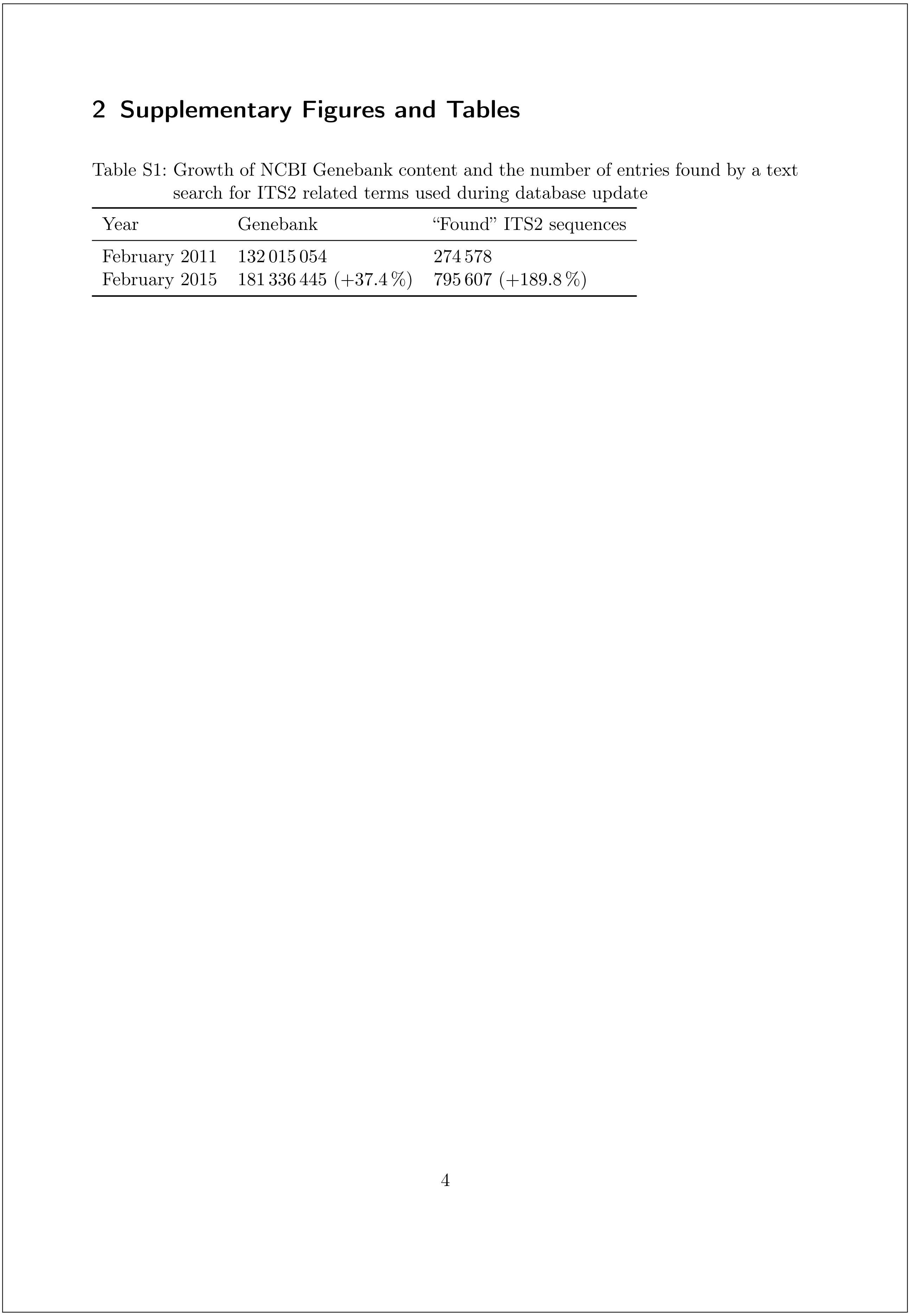
Removed GIs The comparison of the current update with 2011 data set showed 3743 vanished GIs. The reason for that are different: some GIs have been removed, some GIs have been moved into Genebank divisions which are not scanned during database update (e.g. EST), and some GIs have been substituted by newer GIs. To determine for what reason each GI was lost, we checked Genebank for each entry (table S2).

Mapping to BISON database A checklist of all species from the kingdom plantae (taxonomic serial number (TSN): 202422) for all states of the USA were downloaded from Biodiversity Information Serving Our Nation (BISON)(`http://bison.usgs.ornl.gov`) on 5th June 2015. NCBI TaxIDs were assigned on species level via NCBI Taxonomy name/id Status Report Page (`http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi`) on 5th June 2015. The TaxIDs were combined for all states and compared to the list of all species TaxIDs in the ITS2 workbench from 2011 and 2015.

Mapping count information to the tree based on eukaryotic groups according to Adl et al. [1] For 2011 and 2015 data sets a list of GI and GI-derived TaxIDs are generated



using the Perl module `NCBI::Taxonomy` (v0.80) [2]. Due to the groups of Adl et al. [1] belong not necessarily to a single NCBI TaxID, we created a mapping table (table S3). The GI-TaxID files are then used to call the Perl script `generate_adl_mappings.pl`. This generates a bunch of files containing the data set information based on current NCBI Taxonomy. This is later used to create the tree image using iTol [4]. Moreover, the script `generate_adl_mappings.pl` also generates the count information used for the table in the main manuscript.



2 Supplementary Figures and Tables

Table S1: Growth of NCBI Genebank content and the number of entries found by a text search for ITS2 related terms used during database update

Year	Genebank	“Found” ITS2 sequences
February 2011	132 015 054	274 578
February 2015	181 336 445 (+37.4 %)	795 607 (+189.8 %)

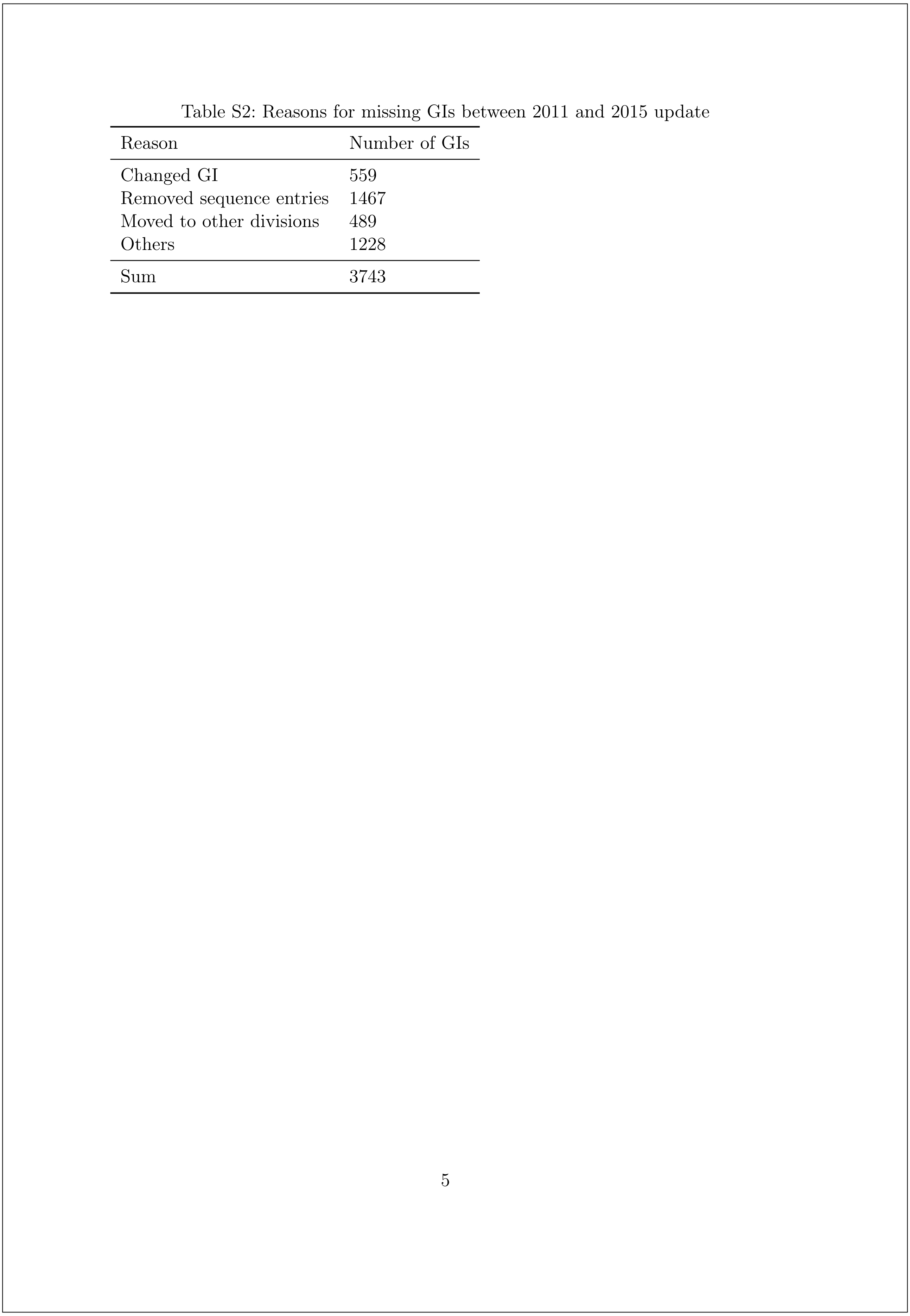
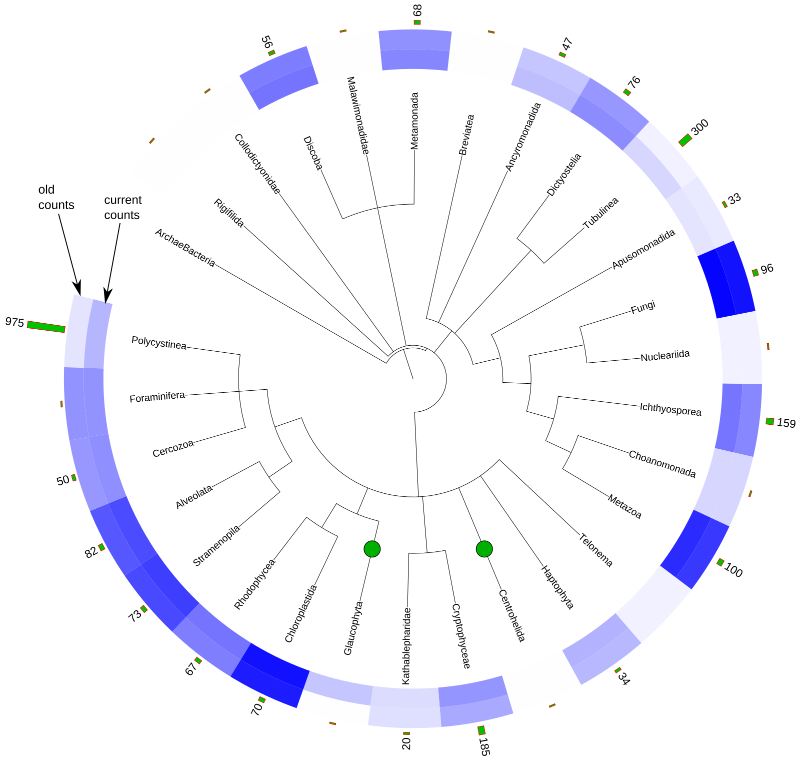


Table S2: Reasons for missing GIs between 2011 and 2015 update

Reason	Number of GIs
Changed GI	559
Removed sequence entries	1467
Moved to other divisions	489
Others	1228
Sum	3743

Figure S1: ITS2 database coverage for the taxonomic tree according to the eukaryotic groups suggested by Adl et al. [1]. The green circles represent groups which are not covered by sequences in 2011 update but covered in current update. The blue arcs indicate the logarithmic absolute counts in the current (inner arcs) and the 2011 update (outer arcs). The green bar plots show the increase in percentage between the old and the current database update.



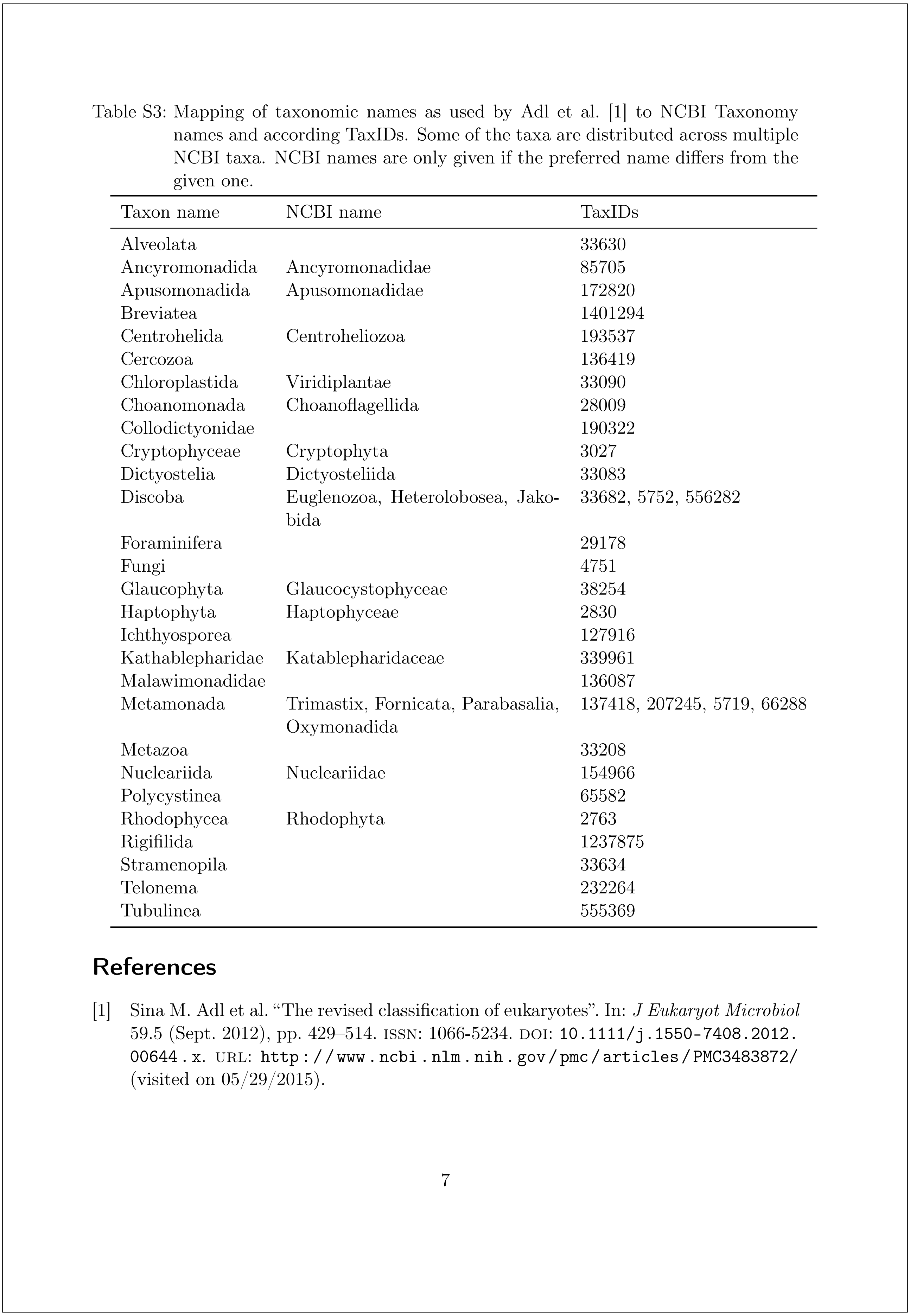
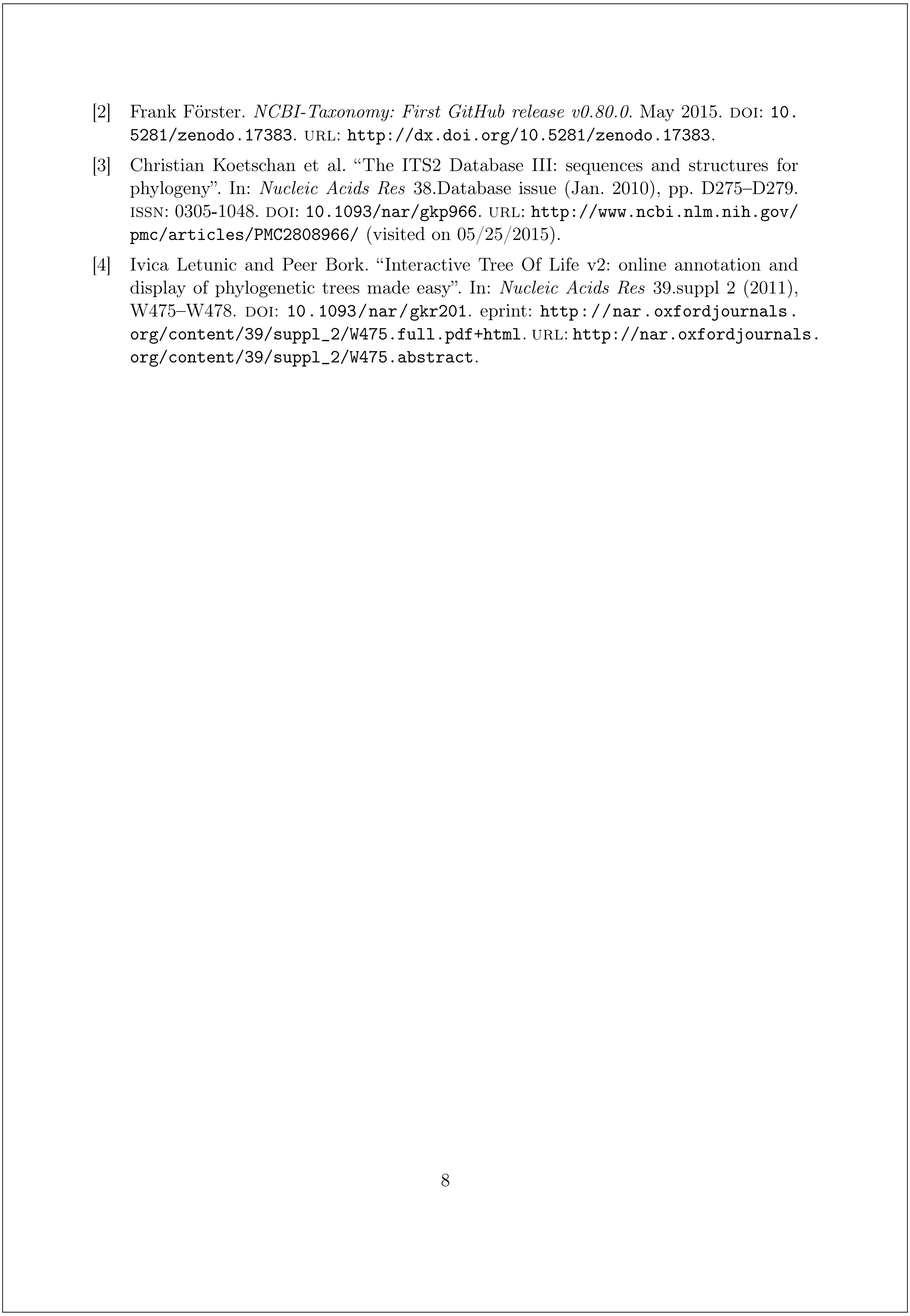


Table S3: Mapping of taxonomic names as used by Adl et al. [1] to NCBI Taxonomy names and according TaxIDs. Some of the taxa are distributed across multiple NCBI taxa. NCBI names are only given if the preferred name differs from the given one.

Taxon name	NCBI name	TaxIDs
Alveolata		33630
Ancyromonadida	Ancyromonadidae	85705
Apusomonadida	Apusomonadidae	172820
Breviatea		1401294
Centrohelida	Centroheliozoa	193537
Cercozoa		136419
Chloroplastida	Viridiplantae	33090
Choanomonada	Choanoflagellida	28009
Collodictyonidae		190322
Cryptophyceae	Cryptophyta	3027
Dictyostelia	Dictyosteliida	33083
Discoba	Euglenozoa, Heterolobosea, Jakobida	33682, 5752, 556282
Foraminifera		29178
Fungi		4751
Glaucophyta	Glaucocystophyceae	38254
Haptophyta	Haptophyceae	2830
Ichthyosporea		127916
Kathablepharidae	Katablepharidaceae	339961
Malawimonadidae		136087
Metamonada	Trimastix, Fornicata, Parabasalia, Oxymonadida	137418, 207245, 5719, 66288
Metazoa		33208
Nucleariida	Nucleariidae	154966
Polycystinea		65582
Rhodophycea	Rhodophyta	2763
Rigifilida		1237875
Stramenopila		33634
Telonema		232264
Tubulinea		555369

References

[1] Sina M. Adl et al. “The revised classification of eukaryotes”. In: *J Eukaryot Microbiol* 59.5 (Sept. 2012), pp. 429–514. ISSN: 1066-5234. DOI: 10.1111/j.1550-7408.2012.00644.x. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3483872/ (visited on 05/29/2015).



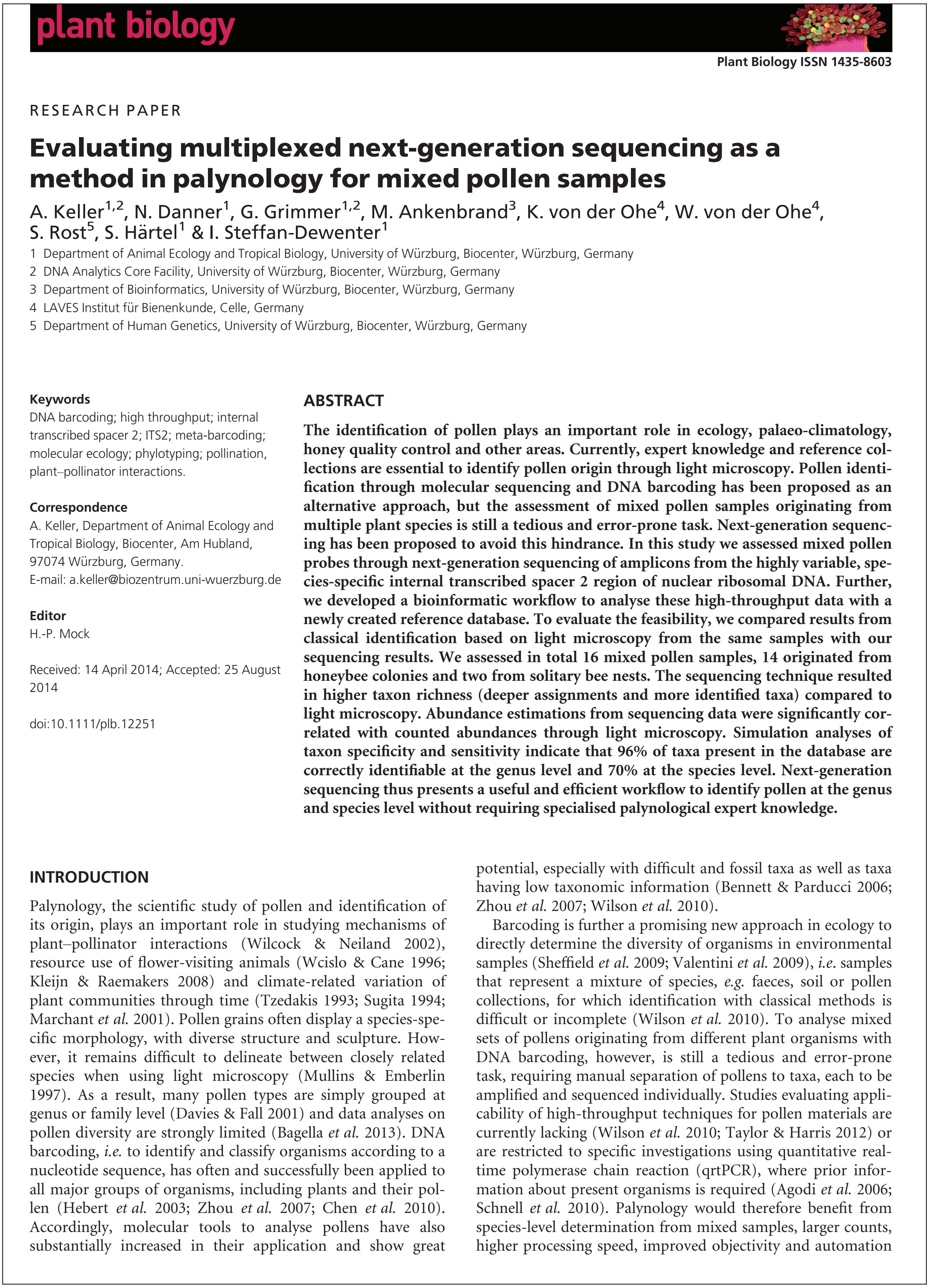
[2] Frank Förster. *NCBI-Taxonomy: First GitHub release v0.80.0.* May 2015. DOI: `10.5281/zenodo.17383`. URL: `http://dx.doi.org/10.5281/zenodo.17383`.

[3] Christian Koetschan et al. "The ITS2 Database III: sequences and structures for phylogeny". In: *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D275–D279. ISSN: 0305-1048. DOI: `10.1093/nar/gkp966`. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808966/` (visited on 05/25/2015).

[4] Ivica Letunic and Peer Bork. "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy". In: *Nucleic Acids Res* 39.suppl 2 (2011), W475–W478. DOI: `10.1093/nar/gkr201`. eprint: `http://nar.oxfordjournals.org/content/39/suppl_2/W475.full.pdf+html`. URL: `http://nar.oxfordjournals.org/content/39/suppl_2/W475.abstract`.

5.2 EVALUATING MULTIPLEXED NEXT-GENERATION SEQUENCING AS A METHOD IN PALYNOLOGY FOR MIXED POLLEN SAMPLES

– published in *Plant Biology* –



plant biology

Plant Biology ISSN 1435-8603

RESEARCH PAPER

Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples

A. Keller[1,2], N. Danner[1], G. Grimmer[1,2], M. Ankenbrand[3], K. von der Ohe[4], W. von der Ohe[4], S. Rost[5], S. Härtel[1] & I. Steffan-Dewenter[1]

1 Department of Animal Ecology and Tropical Biology, University of Würzburg, Biocenter, Würzburg, Germany
2 DNA Analytics Core Facility, University of Würzburg, Biocenter, Würzburg, Germany
3 Department of Bioinformatics, University of Würzburg, Biocenter, Würzburg, Germany
4 LAVES Institut für Bienenkunde, Celle, Germany
5 Department of Human Genetics, University of Würzburg, Biocenter, Würzburg, Germany

Keywords
DNA barcoding; high throughput; internal transcribed spacer 2; ITS2; meta-barcoding; molecular ecology; phylotyping; pollination, plant–pollinator interactions.

Correspondence
A. Keller, Department of Animal Ecology and Tropical Biology, Biocenter, Am Hubland, 97074 Würzburg, Germany.
E-mail: a.keller@biozentrum.uni-wuerzburg.de

Editor
H.-P. Mock

Received: 14 April 2014; Accepted: 25 August 2014

doi:10.1111/plb.12251

ABSTRACT

The identification of pollen plays an important role in ecology, palaeo-climatology, honey quality control and other areas. Currently, expert knowledge and reference collections are essential to identify pollen origin through light microscopy. Pollen identification through molecular sequencing and DNA barcoding has been proposed as an alternative approach, but the assessment of mixed pollen samples originating from multiple plant species is still a tedious and error-prone task. Next-generation sequencing has been proposed to avoid this hindrance. In this study we assessed mixed pollen probes through next-generation sequencing of amplicons from the highly variable, species-specific internal transcribed spacer 2 region of nuclear ribosomal DNA. Further, we developed a bioinformatic workflow to analyse these high-throughput data with a newly created reference database. To evaluate the feasibility, we compared results from classical identification based on light microscopy from the same samples with our sequencing results. We assessed in total 16 mixed pollen samples, 14 originated from honeybee colonies and two from solitary bee nests. The sequencing technique resulted in higher taxon richness (deeper assignments and more identified taxa) compared to light microscopy. Abundance estimations from sequencing data were significantly correlated with counted abundances through light microscopy. Simulation analyses of taxon specificity and sensitivity indicate that 96% of taxa present in the database are correctly identifiable at the genus level and 70% at the species level. Next-generation sequencing thus presents a useful and efficient workflow to identify pollen at the genus and species level without requiring specialised palynological expert knowledge.

INTRODUCTION

Palynology, the scientific study of pollen and identification of its origin, plays an important role in studying mechanisms of plant–pollinator interactions (Wilcock & Neiland 2002), resource use of flower-visiting animals (Wcislo & Cane 1996; Kleijn & Raemakers 2008) and climate-related variation of plant communities through time (Tzedakis 1993; Sugita 1994; Marchant *et al.* 2001). Pollen grains often display a species-specific morphology, with diverse structure and sculpture. However, it remains difficult to delineate between closely related species when using light microscopy (Mullins & Emberlin 1997). As a result, many pollen types are simply grouped at genus or family level (Davies & Fall 2001) and data analyses on pollen diversity are strongly limited (Bagella *et al.* 2013). DNA barcoding, *i.e.* to identify and classify organisms according to a nucleotide sequence, has often and successfully been applied to all major groups of organisms, including plants and their pollen (Hebert *et al.* 2003; Zhou *et al.* 2007; Chen *et al.* 2010). Accordingly, molecular tools to analyse pollens have also substantially increased in their application and show great potential, especially with difficult and fossil taxa as well as taxa having low taxonomic information (Bennett & Parducci 2006; Zhou *et al.* 2007; Wilson *et al.* 2010).

Barcoding is further a promising new approach in ecology to directly determine the diversity of organisms in environmental samples (Sheffield *et al.* 2009; Valentini *et al.* 2009), *i.e.* samples that represent a mixture of species, *e.g.* faeces, soil or pollen collections, for which identification with classical methods is difficult or incomplete (Wilson *et al.* 2010). To analyse mixed sets of pollens originating from different plant organisms with DNA barcoding, however, is still a tedious and error-prone task, requiring manual separation of pollens to taxa, each to be amplified and sequenced individually. Studies evaluating applicability of high-throughput techniques for pollen materials are currently lacking (Wilson *et al.* 2010; Taylor & Harris 2012) or are restricted to specific investigations using quantitative real-time polymerase chain reaction (qrtPCR), where prior information about present organisms is required (Agodi *et al.* 2006; Schnell *et al.* 2010). Palynology would therefore benefit from species-level determination from mixed samples, larger counts, higher processing speed, improved objectivity and automation to be attractive for large-scale studies (Stillman & Flenley 1996). Molecular methods based on high-throughput DNA sequencing could provide the required features to extend and improve classical pollen determination. Valentini *et al.* (2010) proposed next-generation sequencing (NGS) as a suitable method for this task. We agree with this idea, and thus in this study we evaluated performance and reliability of the new sequencing and bioinformatic strategies by directly comparing them with data obtained from light microscopy.

Specifically, we address the following challenges that emerge in DNA barcoding with mixed pollen samples. (i) A laboratory routine has to be defined that can be applied to all major plant clades, requiring universality of amplification priming regions and adequate length to be suitable for next-generation sequencing while holding enough sequence variation to differentiate between species. This routine includes DNA extraction, amplification, sample multiplexing, library preparation, sequencing with high-throughput devices and raw data cleanup. Also, (ii) a mapping algorithm must be developed that adequately maps the obtained sequences in their full length to reference samples, preferably in a hierarchical progression with confidence values for each level of taxonomy. Further, this algorithm has to perform sufficiently well to be able to process high-throughput data on a standard desktop computer and produce results in a reasonable time. (iii) A comprehensive reference database is required to derive the desired taxonomic annotations.

Several genetic marker regions have been proposed for DNA barcoding in plants that match the above requirements, foremost presence and feasibility for amplification in all investigated taxa, as well as low intraspecific but high interspecific variability to succeed in being species-specific (Hebert *et al.* 2003; Zhou *et al.* 2007; Chen *et al.* 2010; Hollingsworth *et al.* 2011). In this study, we use the internal transcribed spacer 2 (ITS2) region, which has been shown to be suitable as a barcode for plants (92.7% successful identifications in 6600 samples (Chen *et al.* 2010; Buchheim *et al.* 2011). Also, the enclosed genetic regions (5.8 S and 28 S) are highly conserved throughout the eukaryotes. Thus a universal primer for the analysis of probes consisting of multiple organisms is applicable, with a low risk of excluding taxa from the amplification (White *et al.* 1990; Keller *et al.* 2009; Chen *et al.* 2010). A further reason for choosing this marker is that a comprehensive ITS2 database already exists (Koetschan *et al.* 2010), enabling preparation of reference sequences suitable for our needs.

We approached the targeted tasks by combining and adapting existing molecular and bioinformatic tools to develop new functionalities for DNA barcoding of pollen samples that consist of multiple taxa. We then evaluated the performance and quality of the molecular and bioinformatic workflow by comparing our results with data from classical light microscopy identification of pollen samples. Further, we tested the applicability for samples with low pollen content and performed computer-based simulations to validate whether the bioinformatic classification pipeline is trustworthy.

MATERIAL AND METHODS

Pollen collection

The honeybee pollen samples were collected in 12 different landscapes in the region around Bayreuth, Germany. The distance between landscapes was at least 3 km, leading to diverse pollen inputs, depending on the surrounding floral resources. In the centre of each landscape we established a honeybee colony (*Apis mellifera carnica* L.) with a pollen trap in front of the hive entrance. Returning foragers had to pass through a 5-mm grid, removing the pollen load from their hind legs. From 21 July 2009 to 12 August 2009, every 1–3 days accumulated pollen loads were removed from the traps and stored as individual samples at −18 °C until the end of the sampling period. Pollen samples were dried at 30 °C for 1 week. Further, to assess variability in resource use of honeybees at one location, samples from three colonies located at the same study site were separately analysed (in the following designated as samples 12a, 12b and 12c). From each of the 14 samples (one per colony), 20% of the collected pollen was randomly taken and mixed for further analyses.

We performed next generation sequencing (NGS) as well as microscopy assessment of the samples. The samples were split into independent aliquots for these separate, blinded analyses. NGS was performed with samples by AK, GG and MA, whereas samples were classified through classical light microscopy by ND with expert guidance from KvO, without knowledge of the other group's results.

Two further pollen samples were obtained from solitary bee nests (*Osmia bicornis* L.) in October 2012 by swabbing the cell walls with cotton buds (Keller *et al.* 2013). In contrast to the relatively pure pollen samples obtained from honeybees, this experiment reflects samples strongly contaminated with nest building material (soil) and faeces, which is challenging to analyse with traditional methods. Solitary bee samples were thus only processed with NGS.

Classical pollen identification

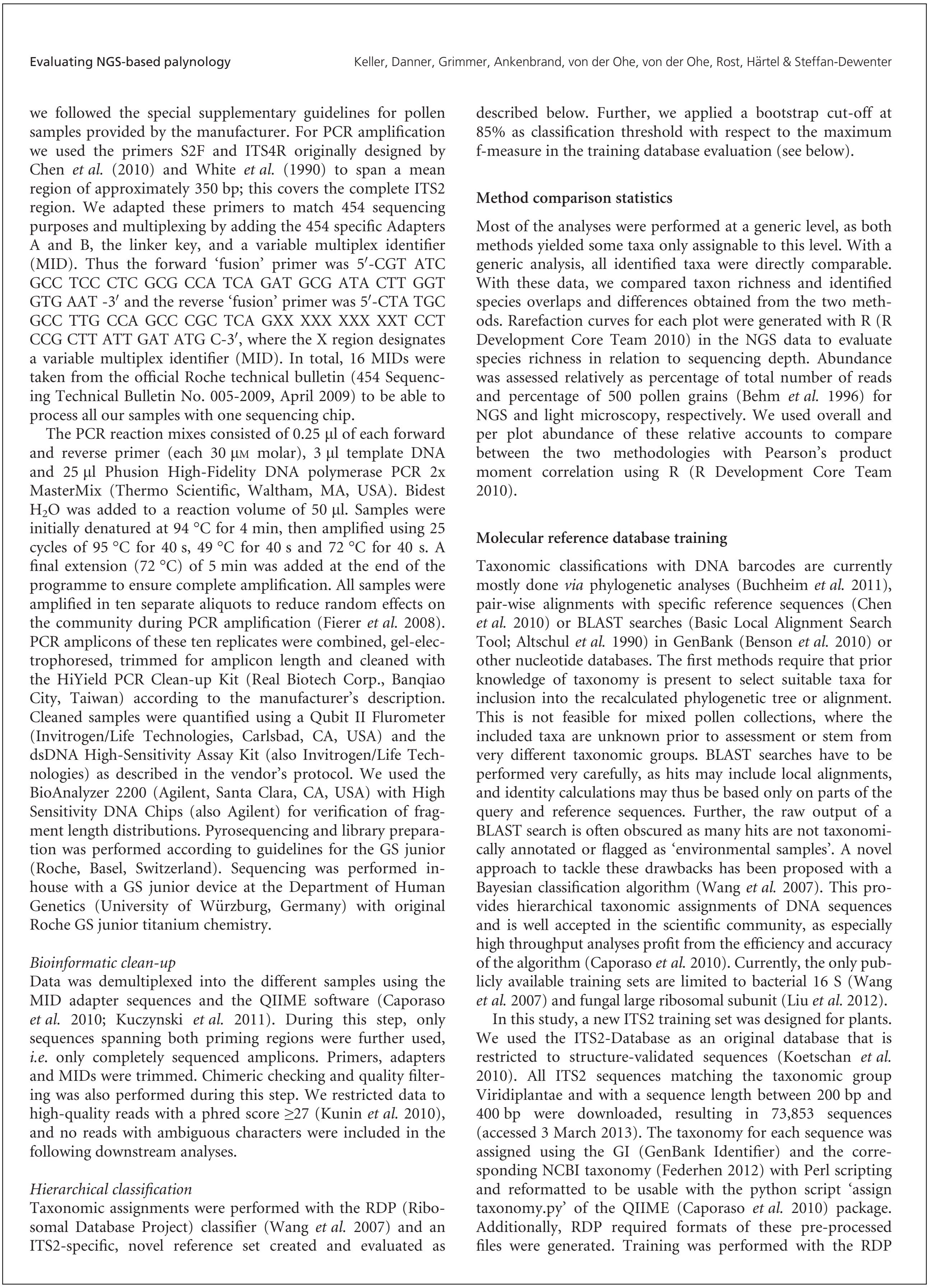
Pollen samples were first analysed using light microscopy in the LAVES Institut für Bienenkunde, Celle, Germany. For microscopic pollen determination, 10 mg pollen loads of each sample were homogenised in 50 ml demineralised water with a magnetic stirrer for 1 h. An aliquot of 15 μl of the solution and 30 μl demineralised water were transferred to a slide, distributed equally over an area of the size of a cover glass and embedded in glycerine:gelatin after complete dehydration, following the method of Behm *et al.* (1996). From each sample, 500 randomly selected pollen grains were determined to genus level, and where possible to species level. Very rarely occurring pollen types were not determined (Behm *et al.* 1996).

Molecular pollen identification

Second pollen identification was done using DNA barcoding of the ITS2 region. The main working steps described below were: DNA extraction, amplification, sequencing, bioinformatic clean-up and taxonomic classification.

DNA extraction, amplification and sequencing

For each sample, 2 g pollens were added to 4 ml bidest H_2O and homogenised with an electronic pestle within a plastic tube. Of this emulsion, 200 μl (~ 50 mg pollens) were taken for the following extraction. We ground the aliquot with a TissueLyser LT (Qiagen, Hilden, Germany) and extracted DNA using the Machery-Nagel (Düren, Germany) NucleoSpin Food Kit;



we followed the special supplementary guidelines for pollen samples provided by the manufacturer. For PCR amplification we used the primers S2F and ITS4R originally designed by Chen *et al.* (2010) and White *et al.* (1990) to span a mean region of approximately 350 bp; this covers the complete ITS2 region. We adapted these primers to match 454 sequencing purposes and multiplexing by adding the 454 specific Adapters A and B, the linker key, and a variable multiplex identifier (MID). Thus the forward 'fusion' primer was 5′-CGT ATC GCC TCC CTC GCG CCA TCA GAT GCG ATA CTT GGT GTG AAT -3′ and the reverse 'fusion' primer was 5′-CTA TGC GCC TTG CCA GCC CGC TCA GXX XXX XXX XXT CCT CCG CTT ATT GAT ATG C-3′, where the X region designates a variable multiplex identifier (MID). In total, 16 MIDs were taken from the official Roche technical bulletin (454 Sequencing Technical Bulletin No. 005-2009, April 2009) to be able to process all our samples with one sequencing chip.

The PCR reaction mixes consisted of 0.25 µl of each forward and reverse primer (each 30 µM molar), 3 µl template DNA and 25 µl Phusion High-Fidelity DNA polymerase PCR 2x MasterMix (Thermo Scientific, Waltham, MA, USA). Bidest H_2O was added to a reaction volume of 50 µl. Samples were initially denatured at 94 °C for 4 min, then amplified using 25 cycles of 95 °C for 40 s, 49 °C for 40 s and 72 °C for 40 s. A final extension (72 °C) of 5 min was added at the end of the programme to ensure complete amplification. All samples were amplified in ten separate aliquots to reduce random effects on the community during PCR amplification (Fierer *et al.* 2008). PCR amplicons of these ten replicates were combined, gel-electrophoresed, trimmed for amplicon length and cleaned with the HiYield PCR Clean-up Kit (Real Biotech Corp., Banqiao City, Taiwan) according to the manufacturer's description. Cleaned samples were quantified using a Qubit II Flurometer (Invitrogen/Life Technologies, Carlsbad, CA, USA) and the dsDNA High-Sensitivity Assay Kit (also Invitrogen/Life Technologies) as described in the vendor's protocol. We used the BioAnalyzer 2200 (Agilent, Santa Clara, CA, USA) with High Sensitivity DNA Chips (also Agilent) for verification of fragment length distributions. Pyrosequencing and library preparation was performed according to guidelines for the GS junior (Roche, Basel, Switzerland). Sequencing was performed in-house with a GS junior device at the Department of Human Genetics (University of Würzburg, Germany) with original Roche GS junior titanium chemistry.

Bioinformatic clean-up

Data was demultiplexed into the different samples using the MID adapter sequences and the QIIME software (Caporaso *et al.* 2010; Kuczynski *et al.* 2011). During this step, only sequences spanning both priming regions were further used, *i.e.* only completely sequenced amplicons. Primers, adapters and MIDs were trimmed. Chimeric checking and quality filtering was also performed during this step. We restricted data to high-quality reads with a phred score $\geq$27 (Kunin *et al.* 2010), and no reads with ambiguous characters were included in the following downstream analyses.

Hierarchical classification

Taxonomic assignments were performed with the RDP (Ribosomal Database Project) classifier (Wang *et al.* 2007) and an ITS2-specific, novel reference set created and evaluated as described below. Further, we applied a bootstrap cut-off at 85% as classification threshold with respect to the maximum f-measure in the training database evaluation (see below).

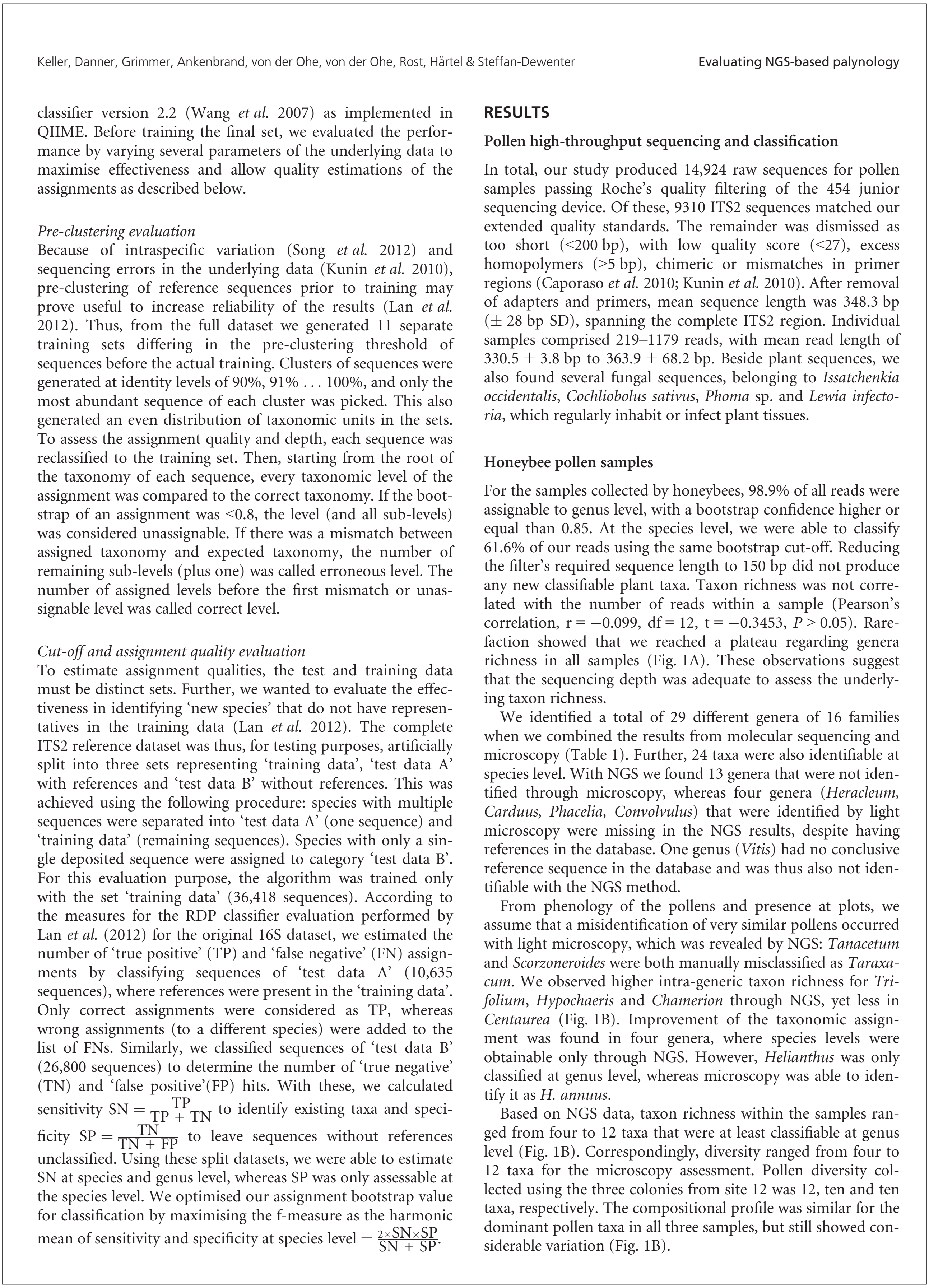
Method comparison statistics

Most of the analyses were performed at a generic level, as both methods yielded some taxa only assignable to this level. With a generic analysis, all identified taxa were directly comparable. With these data, we compared taxon richness and identified species overlaps and differences obtained from the two methods. Rarefaction curves for each plot were generated with R (R Development Core Team 2010) in the NGS data to evaluate species richness in relation to sequencing depth. Abundance was assessed relatively as percentage of total number of reads and percentage of 500 pollen grains (Behm *et al.* 1996) for NGS and light microscopy, respectively. We used overall and per plot abundance of these relative accounts to compare between the two methodologies with Pearson's product moment correlation using R (R Development Core Team 2010).

Molecular reference database training

Taxonomic classifications with DNA barcodes are currently mostly done *via* phylogenetic analyses (Buchheim *et al.* 2011), pair-wise alignments with specific reference sequences (Chen *et al.* 2010) or BLAST searches (Basic Local Alignment Search Tool; Altschul *et al.* 1990) in GenBank (Benson *et al.* 2010) or other nucleotide databases. The first methods require that prior knowledge of taxonomy is present to select suitable taxa for inclusion into the recalculated phylogenetic tree or alignment. This is not feasible for mixed pollen collections, where the included taxa are unknown prior to assessment or stem from very different taxonomic groups. BLAST searches have to be performed very carefully, as hits may include local alignments, and identity calculations may thus be based only on parts of the query and reference sequences. Further, the raw output of a BLAST search is often obscured as many hits are not taxonomically annotated or flagged as 'environmental samples'. A novel approach to tackle these drawbacks has been proposed with a Bayesian classification algorithm (Wang *et al.* 2007). This provides hierarchical taxonomic assignments of DNA sequences and is well accepted in the scientific community, as especially high throughput analyses profit from the efficiency and accuracy of the algorithm (Caporaso *et al.* 2010). Currently, the only publicly available training sets are limited to bacterial 16 S (Wang *et al.* 2007) and fungal large ribosomal subunit (Liu *et al.* 2012).

In this study, a new ITS2 training set was designed for plants. We used the ITS2-Database as an original database that is restricted to structure-validated sequences (Koetschan *et al.* 2010). All ITS2 sequences matching the taxonomic group Viridiplantae and with a sequence length between 200 bp and 400 bp were downloaded, resulting in 73,853 sequences (accessed 3 March 2013). The taxonomy for each sequence was assigned using the GI (GenBank Identifier) and the corresponding NCBI taxonomy (Federhen 2012) with Perl scripting and reformatted to be usable with the python script 'assign taxonomy.py' of the QIIME (Caporaso *et al.* 2010) package. Additionally, RDP required formats of these pre-processed files were generated. Training was performed with the RDP



classifier version 2.2 (Wang *et al.* 2007) as implemented in QIIME. Before training the final set, we evaluated the performance by varying several parameters of the underlying data to maximise effectiveness and allow quality estimations of the assignments as described below.

Pre-clustering evaluation

Because of intraspecific variation (Song *et al.* 2012) and sequencing errors in the underlying data (Kunin *et al.* 2010), pre-clustering of reference sequences prior to training may prove useful to increase reliability of the results (Lan *et al.* 2012). Thus, from the full dataset we generated 11 separate training sets differing in the pre-clustering threshold of sequences before the actual training. Clusters of sequences were generated at identity levels of 90%, 91% . . . 100%, and only the most abundant sequence of each cluster was picked. This also generated an even distribution of taxonomic units in the sets. To assess the assignment quality and depth, each sequence was reclassified to the training set. Then, starting from the root of the taxonomy of each sequence, every taxonomic level of the assignment was compared to the correct taxonomy. If the bootstrap of an assignment was <0.8, the level (and all sub-levels) was considered unassignable. If there was a mismatch between assigned taxonomy and expected taxonomy, the number of remaining sub-levels (plus one) was called erroneous level. The number of assigned levels before the first mismatch or unassignable level was called correct level.

Cut-off and assignment quality evaluation

To estimate assignment qualities, the test and training data must be distinct sets. Further, we wanted to evaluate the effectiveness in identifying 'new species' that do not have representatives in the training data (Lan *et al.* 2012). The complete ITS2 reference dataset was thus, for testing purposes, artificially split into three sets representing 'training data', 'test data A' with references and 'test data B' without references. This was achieved using the following procedure: species with multiple sequences were separated into 'test data A' (one sequence) and 'training data' (remaining sequences). Species with only a single deposited sequence were assigned to category 'test data B'. For this evaluation purpose, the algorithm was trained only with the set 'training data' (36,418 sequences). According to the measures for the RDP classifier evaluation performed by Lan *et al.* (2012) for the original 16S dataset, we estimated the number of 'true positive' (TP) and 'false negative' (FN) assignments by classifying sequences of 'test data A' (10,635 sequences), where references were present in the 'training data'. Only correct assignments were considered as TP, whereas wrong assignments (to a different species) were added to the list of FNs. Similarly, we classified sequences of 'test data B' (26,800 sequences) to determine the number of 'true negative' (TN) and 'false positive'(FP) hits. With these, we calculated sensitivity $\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}}$ to identify existing taxa and specificity $\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$ to leave sequences without references unclassified. Using these split datasets, we were able to estimate SN at species and genus level, whereas SP was only assessable at the species level. We optimised our assignment bootstrap value for classification by maximising the f-measure as the harmonic mean of sensitivity and specificity at species level $= \frac{2\times\mathrm{SN}\times\mathrm{SP}}{\mathrm{SN} + \mathrm{SP}}$.

RESULTS

Pollen high-throughput sequencing and classification

In total, our study produced 14,924 raw sequences for pollen samples passing Roche's quality filtering of the 454 junior sequencing device. Of these, 9310 ITS2 sequences matched our extended quality standards. The remainder was dismissed as too short (<200 bp), with low quality score (<27), excess homopolymers (>5 bp), chimeric or mismatches in primer regions (Caporaso *et al.* 2010; Kunin *et al.* 2010). After removal of adapters and primers, mean sequence length was 348.3 bp (± 28 bp SD), spanning the complete ITS2 region. Individual samples comprised 219–1179 reads, with mean read length of 330.5 ± 3.8 bp to 363.9 ± 68.2 bp. Beside plant sequences, we also found several fungal sequences, belonging to *Issatchenkia occidentalis*, *Cochliobolus sativus*, *Phoma* sp. and *Lewia infectoria*, which regularly inhabit or infect plant tissues.

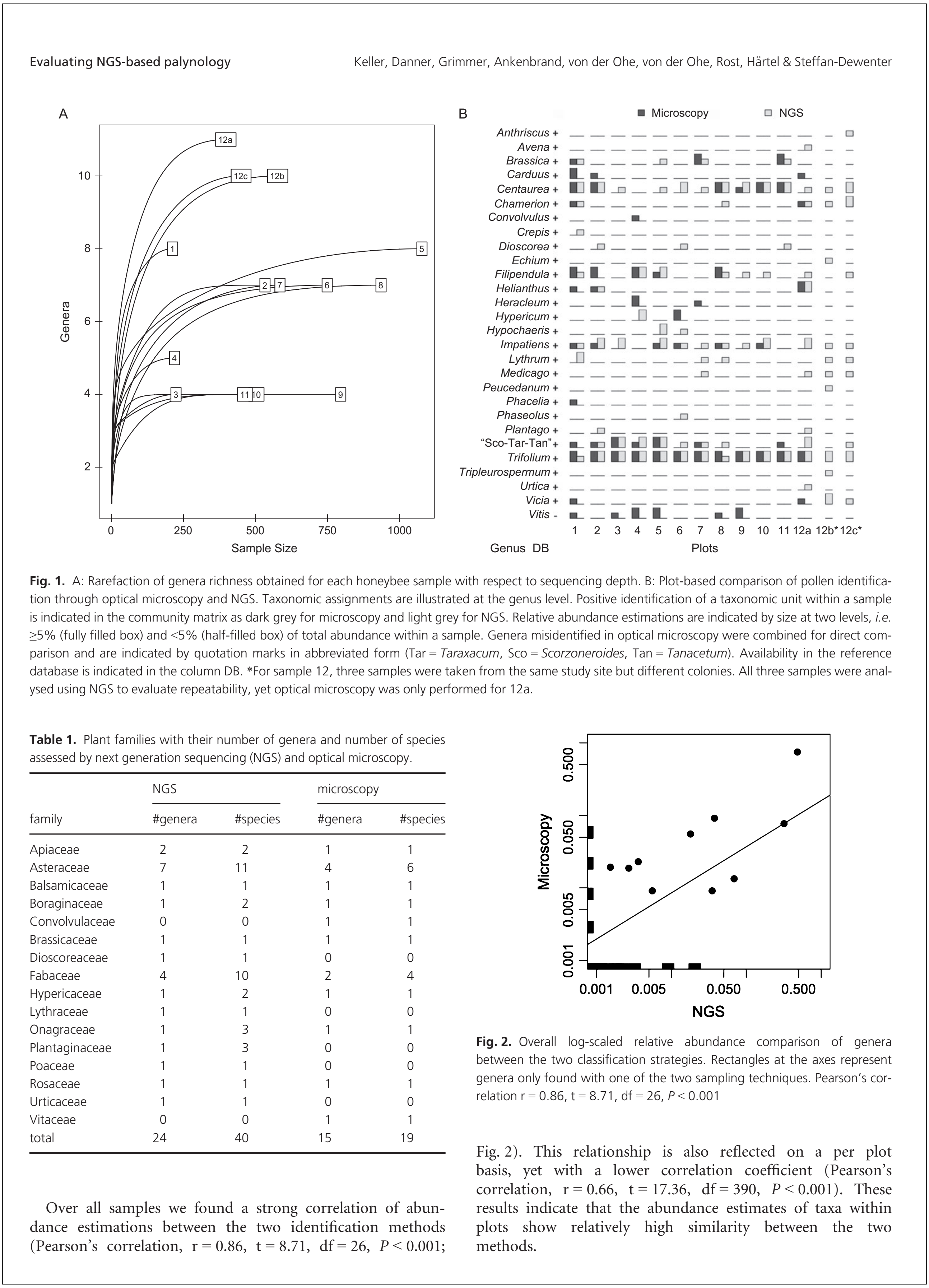
Honeybee pollen samples

For the samples collected by honeybees, 98.9% of all reads were assignable to genus level, with a bootstrap confidence higher or equal than 0.85. At the species level, we were able to classify 61.6% of our reads using the same bootstrap cut-off. Reducing the filter's required sequence length to 150 bp did not produce any new classifiable plant taxa. Taxon richness was not correlated with the number of reads within a sample (Pearson's correlation, $r = -0.099$, $df = 12$, $t = -0.3453$, $P > 0.05$). Rarefaction showed that we reached a plateau regarding genera richness in all samples (Fig. 1A). These observations suggest that the sequencing depth was adequate to assess the underlying taxon richness.

We identified a total of 29 different genera of 16 families when we combined the results from molecular sequencing and microscopy (Table 1). Further, 24 taxa were also identifiable at species level. With NGS we found 13 genera that were not identified through microscopy, whereas four genera (*Heracleum, Carduus, Phacelia, Convolvulus*) that were identified by light microscopy were missing in the NGS results, despite having references in the database. One genus (*Vitis*) had no conclusive reference sequence in the database and was thus also not identifiable with the NGS method.

From phenology of the pollens and presence at plots, we assume that a misidentification of very similar pollens occurred with light microscopy, which was revealed by NGS: *Tanacetum* and *Scorzoneroides* were both manually misclassified as *Taraxacum*. We observed higher intra-generic taxon richness for *Trifolium*, *Hypochaeris* and *Chamerion* through NGS, yet less in *Centaurea* (Fig. 1B). Improvement of the taxonomic assignment was found in four genera, where species levels were obtainable only through NGS. However, *Helianthus* was only classified at genus level, whereas microscopy was able to identify it as *H. annuus*.

Based on NGS data, taxon richness within the samples ranged from four to 12 taxa that were at least classifiable at genus level (Fig. 1B). Correspondingly, diversity ranged from four to 12 taxa for the microscopy assessment. Pollen diversity collected using the three colonies from site 12 was 12, ten and ten taxa, respectively. The compositional profile was similar for the dominant pollen taxa in all three samples, but still showed considerable variation (Fig. 1B).



A
Genera
2
4
6
8
10
0
250
500
750
1000
Sample Size
12a
12c
12b
1
5
2
7
6
8
4
3
11
10
9
B
Microscopy
NGS
Anthriscus +
Avena +
Brassica +
Carduus +
Centaurea +
Chamerion +
Convolvulus +
Crepis +
Dioscorea +
Echium +
Filipendula +
Helianthus +
Heracleum +
Hypericum +
Hypochaeris +
Impatiens +
Lythrum +
Medicago +
Peucedanum +
Phacelia +
Phaseolus +
Plantago +
"Sco-Tar-Tan"+
Trifolium +
Tripleurospermum +
Urtica +
Vicia +
Vitis -
Genus DB
1 2 3 4 5 6 7 8 9 10 11 12a 12b* 12c*
Plots

Fig. 1. A: Rarefaction of genera richness obtained for each honeybee sample with respect to sequencing depth. B: Plot-based comparison of pollen identification through optical microscopy and NGS. Taxonomic assignments are illustrated at the genus level. Positive identification of a taxonomic unit within a sample is indicated in the community matrix as dark grey for microscopy and light grey for NGS. Relative abundance estimations are indicated by size at two levels, *i.e.* ≥5% (fully filled box) and <5% (half-filled box) of total abundance within a sample. Genera misidentified in optical microscopy were combined for direct comparison and are indicated by quotation marks in abbreviated form (Tar = *Taraxacum*, Sco = *Scorzoneroides*, Tan = *Tanacetum*). Availability in the reference database is indicated in the column DB. *For sample 12, three samples were taken from the same study site but different colonies. All three samples were analysed using NGS to evaluate repeatability, yet optical microscopy was only performed for 12a.

Table 1. Plant families with their number of genera and number of species assessed by next generation sequencing (NGS) and optical microscopy.

	NGS		microscopy	
family	#genera	#species	#genera	#species
Apiaceae	2	2	1	1
Asteraceae	7	11	4	6
Balsamicaceae	1	1	1	1
Boraginaceae	1	2	1	1
Convolvulaceae	0	0	1	1
Brassicaceae	1	1	1	1
Dioscoreaceae	1	1	0	0
Fabaceae	4	10	2	4
Hypericaceae	1	2	1	1
Lythraceae	1	1	0	0
Onagraceae	1	3	1	1
Plantaginaceae	1	3	0	0
Poaceae	1	1	0	0
Rosaceae	1	1	1	1
Urticaceae	1	1	0	0
Vitaceae	0	0	1	1
total	24	40	15	19

Fig. 2. Overall log-scaled relative abundance comparison of genera between the two classification strategies. Rectangles at the axes represent genera only found with one of the two sampling techniques. Pearson's correlation r = 0.86, t = 8.71, df = 26, $P < 0.001$

Over all samples we found a strong correlation of abundance estimations between the two identification methods (Pearson's correlation, r = 0.86, t = 8.71, df = 26, $P < 0.001$; Fig. 2). This relationship is also reflected on a per plot basis, yet with a lower correlation coefficient (Pearson's correlation, r = 0.66, t = 17.36, df = 390, $P < 0.001$). These results indicate that the abundance estimates of taxa within plots show relatively high similarity between the two methods.

Pollens in solitary bee nests

Pollen samples from both solitary bee nests were successfully processed, with 100% of reads identifiable at genus level despite high contamination of the samples with nesting material and faeces. Both samples harboured *Brassica* sp. and *Dioscorea* sp. pollen, the latter most likely *Dioscorea* (*Tamus*) *communis* as the only representative of the Dioscoreaceae present in the sampling region.

Molecular reference database training

Pre-clustering of data prior to training of the RDP classifier did not improve the overall performance of classifications (Fig. 3). This was the case both for depth of the assignment as well as the mean number of incorrectly assigned levels, which, respectively, increase and decrease with higher pre-clustering thresholds. We thus used a cut-off at 100% sequence identity, which equals unique sequences, for the final training set. With that, of the 73,853 tested database sequences, 55,028 were positively identifiable at species and a further 10,518 at genus level. Surprisingly, 6104 sequences were assignable only to phylum level; they likely represent contamination in the reference database.

Regarding determination of the optimal cut-off threshold, specificity and sensitivity of the novel/known classifications are shown with their dependency of the bootstrap (Fig. 4). The best classification by means of f-measures is achieved with a bootstrap cut-off of 0.85. Both specificity and sensitivity at this threshold for species level were approximately 70%. At genus level, sensitivity to correctly identify a genus increased to 96%. We thus recommend this threshold when using the RDP classifier with the generated training data.

Currently, all sequences in the reference dataset accumulate to 37,435 different plant species and 6162 genera according to NCBI taxonomy (Federhen 2012). The complete reference dataset is available for download and public use at http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

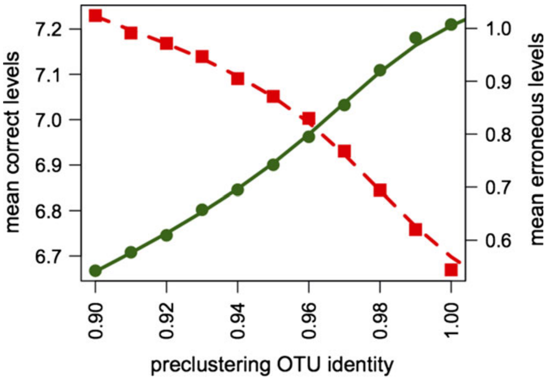


Fig. 3. Pre-clustering evaluation: Starting from the root of the taxonomy of each sequence, every taxonomic level of the assignment was compared to its correct lineage. The overall mean of correct assignments according to the different pre-clustering levels is presented as dots in the figure (left scale). Similarly, each sequence was tested for erroneous levels of classification with means displayed as squares and the scale on the right side.

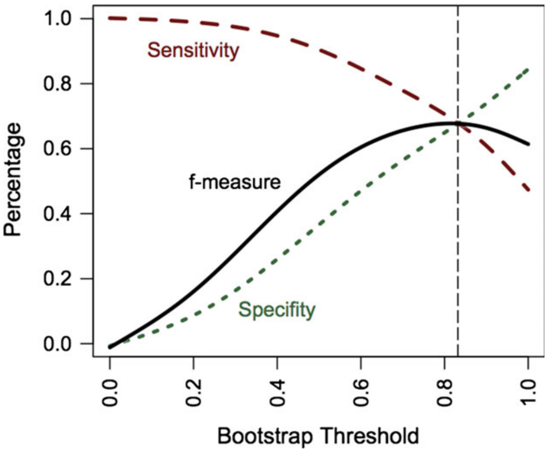


Fig. 4. Dependence of sensitivity and specificity by the bootstrap threshold. Sensitivity to identify at species level is illustrated with a single-dashed line. Specificity is displayed as a dotted line. The harmonic mean of both species-level measures is displayed as a solid black curve, maximised at approximately 0.85 as the suggested optimal classification threshold.

DISCUSSION

The demand for methods to identify pollen samples at a high-throughput level is increasing for many applications in ecology and paleo-climatology (Bennett & Parducci 2006; Zhou *et al.* 2007; Sheffield *et al.* 2009; Valentini *et al.* 2009; Wilson *et al.* 2010; Taylor & Harris 2012). DNA barcoding is a frequently and successfully applied method, yet pollens of mixed samples originating from more than one source are currently not assessable through standard methods. Valentini *et al.* (2010) proposed that NGS may counter this deficiency, *i.e.* to investigate such mixed samples by identifying all included plant organisms together, without manual separation. The goals of this study were thus to develop, and moreover evaluate, a molecular laboratory procedure and bioinformatic analysis for such a task. The complete workflow was applied to pollen samples from two different studies (in total 16 samples). The resulting gene sequences allowed us to successfully identify taxon richness and abundance of the underlying samples. The resulting taxonomic resolution is similar or better than results from classical light microscopy. Details of the performance of each individual step of the workflow and the resulting methodological and biological relevance are discussed below.

High-throughput pollen sequencing

In general, our laboratory workflow was suitable for processing mixed pollen probes through NGS. However, quality filtering according to our rigorous restrictions reduced the obtained sequences from approximately 15,000 sequences to 10,000. Most of them were removed due to failure to include both primer regions and/or multiplex identifier due to low quality scores towards the end of sequences or short read lengths (Caporaso *et al.* 2010). The former indicates that a large proportion of reads was not fully sequenced with sufficient quality, whereas the latter shows that the primers also amplified shorter fragments than the intended plant ITS2 region. Not fully sequenced reads are a technical issue that is regularly improved by increasing read length and quality through new generations of sequencing devices and chemistry (Metzker 2009). Improvements can also be expected by applying paired-end strategies, as quality near the ends will increase, or using technologies with general lower sequencing error rates. Shorter, fully sequenced sequences are project-specific problems, but are also expected: as a drawback of universal primers, they will also amplify fungal ITS2 (White *et al.* 1990) ranging from approximately 100 to 250 bp, and even other eukaryotic protists with far shorter ITS2 regions (Keller *et al.* 2009). Further, the existence of non-functional pseudo-genes is known (Harpke & Peterson 2008). Thus studies investigating plant ITS2 sequences should account for a sufficient overhead of estimated sequences per sample during project design related to sequencing technology and potential contamination from unwanted organisms (Parameswaran *et al.* 2007). The remaining high-quality reads showed a high proportion of classifiable sequences (~99%), whereas reduction of the minimum sequence length had no impact on plant species diversity. Both observations suggest that the filters are adequate to concentrate on the data of interest, *i.e.* plant sequences.

Classification pipeline

To be able to use the RDP classifier (Wang *et al.* 2007) for taxonomic assignments with plants and with the ITS2 marker, we re-trained the algorithm with structurally verified sequences obtained from the ITS2 database (Koetschan *et al.* 2010). The underlying dataset incorporates more than 70,000 different plant sequences and represents a cross-section throughout the Viridiplantae. Sequences originate from all biogeographic regions of the world since the primary database is GenBank (Benson *et al.* 2010). Currently, all sequences in the reference dataset represent 37,435 different plant species and 6162 genera according to NCBI taxonomy (Federhen 2012). Exemplarily for the data analysed in this study, the dataset covers 79% of all vascular plant genera and 54% of species known to exist within the Federal state of Bavaria, Germany, where our samples were obtained (comprehensive plant database http://www.bayernflora.de, accessed 6 November 2013; Staatliche Naturwissenschaftliche Sammlungen Bayern 2013). As 99% of reads were classifiable to genus level and only one genus (*Vitis*) of the assessed 29 genera in total was missing in the reference database, most of the abundant and bee relevant plant genera seem to be included. Further, the classifier's dataset is updateable to match the constantly increasing number of sequences deposited in GenBank and the ITS2 database in the future (Wang *et al.* 2007).

In the computational evaluation of database and classifier for an ITS2 dataset, we obtained values comparable to those of existing datasets published for bacteria (Wang *et al.* 2007) and fungi (Liu *et al.* 2012). Taxonomic classifications performed best regarding sensitivity, *i.e.* to identify taxa existing in the database, and specificity, *i.e.* to restrain from classifying organisms without references, at a bootstrap threshold level of approximately 0.85 (Lan *et al.* 2012). Species- and genus-level sensitivity to correctly identify sequences with this bootstrap were 70% and 96%, respectively. This is similar to the classifier's preferred level used to classify microbial organisms (0.80; Wang *et al.* 2007; Lan *et al.* 2012). From a technical perspective, it is thus valid to also apply the classification algorithm for ITS2 sequences of plants.

Comparison of assessment methods

Using NGS, we were clearly able to improve palynology diversity assessments in comparison with traditional optical microscopy. This appears in novel taxa that were identified, as well as improvement of classification of taxa and better possibilities to distinguish species within a genus. Further, some misidentifications of pollen through microscopy were revealed that were caused by very similar morphological appearance of closely related species. Also, molecular assessments were successful for solitary bee nest samples, where swabs included pollens as well as contaminating material. Sequencing assessments were repeatable, identifying similar diversity in samples obtained from different bee colonies placed within the same landscape.

However, using the high-throughput approach we also encountered limitations, which are partly related to the data used for training of the classifier. Regarding the Vitaceae, the ITS2 database is currently lacking acceptable reference sequences. We validated the only existing sequence, which was very short (~200 bp) and derived from a whole genome shotgun sequencing study (assembled sequence from short length reads, GenBank ID: AM462492.2; Velasco *et al.* 2007). Due to intra-genomic variation of the ITS2 (Song *et al.* 2012), we assume the assembly yielded a consensus, stacked ITS2 sequence, not usable for barcoding purposes or that a non-ITS2 region was falsely identified as such by the ITS2 database annotation algorithm (Keller *et al.* 2009). We therefore dismissed the sequence as missing within the reference database. In general, taxa missing or with inadequate sequences in the underlying database are not identifiable. As shown exemplarily for the geographic region of Bavaria, 22% of known plant genera are missing, and thus the current coverage is far from complete (Staatliche Naturwissenschaftliche Sammlungen Bayern 2013). Also, valid sequences with wrong taxonomic annotations may lead to mis-training of the classification model regarding the respective taxa (Bridge *et al.* 2003). This is highlighted by a proportion of sequences re-classified in the evaluation to a different phylum, suggesting wrong taxonomic annotation of GenBank database sequences. To address limitations of the underlying database (missing or misclassified sequences) in a given research question, we suggest that applied studies should also consider reviewing one cross-section pool of all samples in parallel through optical means to verify the overall richness of taxa relevant for the study. This will also maintain comparability between studies applying traditional and molecular approaches. Despite these database-specific drawbacks, the classifier produced taxonomic assignments that are congruent with light microscopy, and thus corroborating the positive technical evaluation of the pipeline above with a direct comparison of biological data.

Abundance estimations for both methods showed a strong correlation, suggesting that abundance estimates based on high-throughput sequencing regarding high or low sequence frequency of taxa within the sample are valid. In our study, we took care to reduce amplification biases through PCR with ten aliquots of each sample simultaneously (typical in microbiota studies: three, Fierer *et al.* 2008) and a low number of amplification cycles (Suzuki & Giovannoni 1996). Nevertheless, abundances retained from PCR-amplified DNA samples must be regarded critically, as amplification biases through priming preference of specific taxonomic groups, random effects and the exponential nature of the amplification process cannot be excluded (Suzuki & Giovannoni 1996; Spooner 2009). Abundances are thus likely better interpreted as categorical (*e.g.* high abundance, low abundance) than with linear association. With the advent of increased sequencing throughput and third-generation single molecule sequencers without need for amplification (Metzker 2009; Roberts *et al.* 2013), improved abundance estimations from sequencing are likely in the near future.

Cost per sample was almost equal for both applied methods when considering time and consumables. As the trend of sequencing technologies moves rapidly toward higher throughput and resulting multiplexing possibilities (Metzker 2009; Kozich *et al.* 2013), we expect price efficiency per sample with NGS to outpace optical assessments in the near future.

Fields of application

Various applications arise for the proposed method. These include studies of pollen material of various origin, including plants themselves, pollinators, soil samples and wind collections. The results of such assessments are of great importance in identifying the diversity and specialisation of plant–pollinator interaction networks (Bosch *et al.* 2009) and also in supporting agricultural and ecological management decisions (*e.g.* Girard *et al.* 2012; Odoux *et al.* 2012). Further, paleo-ecological and climate change-associated studies investigating fossil pollens may also profit (Bennett & Parducci 2006).

Special attention is currently required in quality control of honeybee products, including the geographic origin, correct labelling of different varieties based on the used floral resources and detection of contamination from genetically modified (GM) crops (Picard-Nizou *et al.* 1995; Hemmer 1997). As pollen is naturally incorporated into honey and protocols to isolate pollens are common usage (Sowunmi 1976), high-throughput sequencing and classification may make a large contribution to this endeavour by facilitating the analytical process and inclusion of references from plant taxa throughout the world (Sowunmi 1976; Ruoff *et al.* 2007).

Furthermore, the methodology may be equivalently applied to other questions not only related to pollens. Other target samples are naturally occurring communities of plants, (*e.g.* green algae) or artificially mixed probes of plant tissue fragments (Schlumbaum *et al.* 2008). As the primers used in this study also efficiently amplify fungal ITS2 sequences, ancillary information is automatically gained about this group, including pathogens as *Ascosphaera* spp. that may be present in collected pollen samples and vectored through harvesting flights of worker bees (Gilliam 1990; White *et al.* 1990).

CONCLUSIONS

Expert knowledge is essential to adequately identify pollens through traditional light microscopy, while taxonomic expertise is also often restricted to specific plant groups or geographic regions. Further, mixed samples of pollens from several plant origins present a problem in current palynology. With this study we evaluated NGS to approach pollen assessments through molecular techniques including their bioinformatic analysis. The analytical pipeline is designed for high-throughput data, but also adaptable to single sequences. It is a useful technique, broadening the assessment capabilities from expert labs to all work groups with access to standard molecular laboratory equipment. Further, our results show that this assessment method improves the standard technique with regard to taxonomic depth, overall diversity and rectifying misidentifications.

ACKNOWLEDGEMENTS

We greatly appreciate the help of staff from the Department of Human Genetics (University of Würzburg, Germany), especially C. Müller-Reible and T. Haaf for placing the 454 device at our disposal for this study. We also thank F. Förster (Department of Bioinformatics, University of Würzburg, Germany) for providing a useful taxonomy Perl module for generation of the training sets, and two anonymous reviewers for their valuable comments.

DATA ACCESSIBILITY

Sequences have been deposited at the ENA:SRA (https://www.ebi.ac.uk/ena) and are accessible under study accession number PRJEB5016. The used training set alongside installation and application notes, is available for download at http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

REFERENCES

Agodi A., Barchitta M., Grillo A., Sciacca S. (2006) Detection of genetically modified DNA sequences in milk from the Italian market. *International Journal of Hygiene and Environmental Health*, **209**, 81–88.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Bagella S., Satta A., Floris I., Caria M.C., Rossetti I., Podani J. (2013) Effects of plant community composition and flowering phenology on honeybee foraging in Mediterranean sylvo-pastoral systems. *Applied Vegetation Science*, **16**, 689–697.

Behm F., von der Ohe K., Henrich W. (1996) Zuverlässigkeit der Pollenanalyse von Honig: bestimmung der Pollenhäufigkeit. *Deutsche Lebensmittel-Rundschau*, **92**, 183–188.

Bennett K.D., Parducci L. (2006) DNA from pollen: principles and potential. *The Holocene*, **16**, 1031–1034.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2010) GenBank. *Nucleic Acids Research*, **38** (Suppl. 1), D46–D51.

Bosch J., Martín González A.M., Rodrigo A., Navarro D. (2009) Plant–pollinator networks: adding the pollinator's perspective. *Ecology Letters*, **12**, 409–419.

Bridge P.D., Roberts P.J., Spooner B.M., Panchal G. (2003) On the unreliability of published DNA sequences. *New Phytologist*, **160**, 43–48.

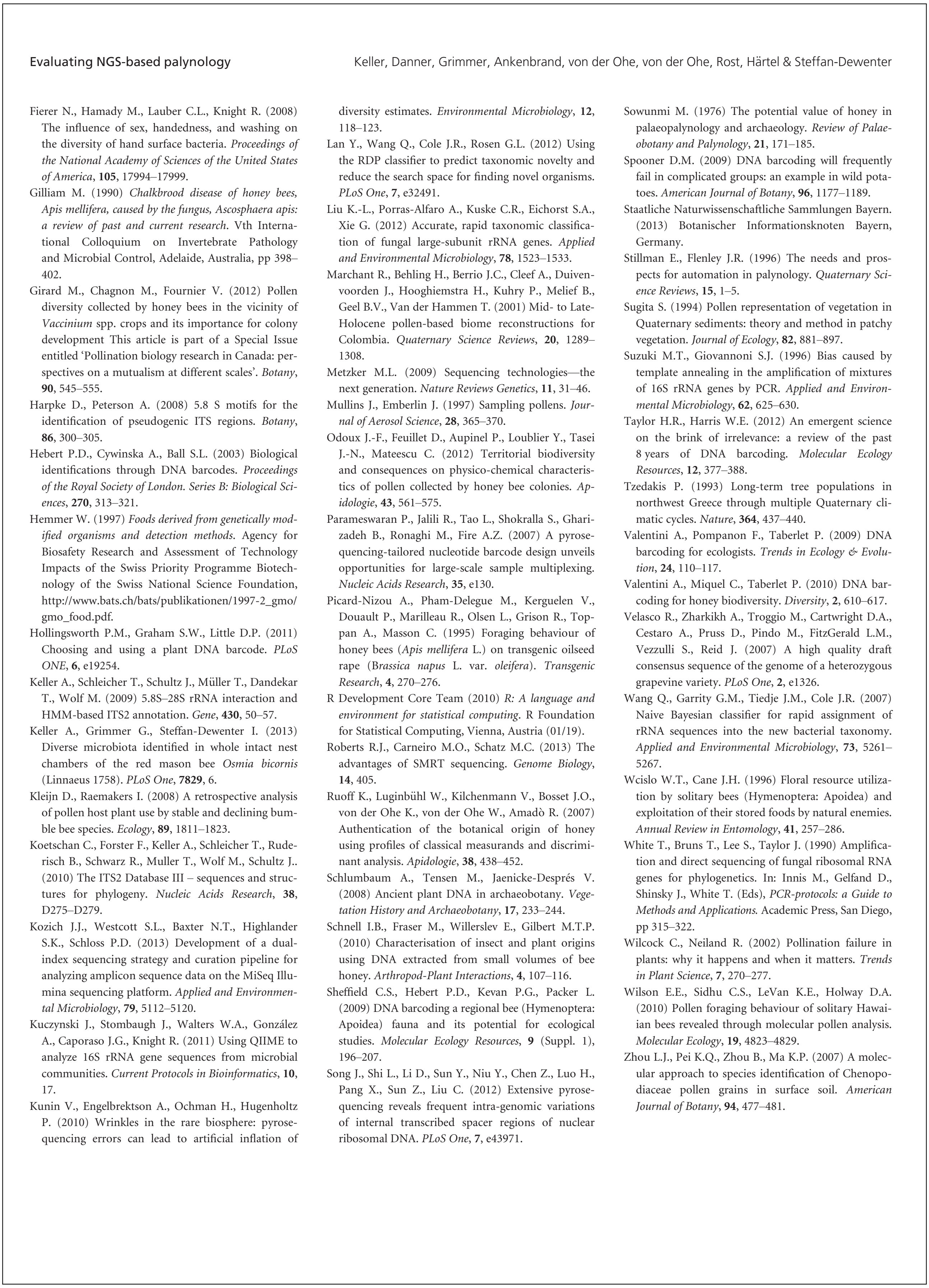
Buchheim M., Keller A., Koetschan C., Forster F., Merget B., Wolf M. (2011) Internal Transcribed Spacer 2 (nu ITS2 rRNA) sequence–structure phylogenetics: towards an automated reconstruction of the green algal tree of life. *PLoS One*, **6**.

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

Chen S., Yao H., Han J., Liu C., Song J., Shi L., Zhu Y., Ma X., Gao T., Pang X. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, **5**, e8613.

Davies C.P., Fall P.L. (2001) Modern pollen precipitation from an elevational transect in central Jordan and its relationship to vegetation. *Journal of Biogeography*, **28**, 1195–1210.

Federhen S. (2012) The NCBI taxonomy database. *Nucleic Acids Research*, **40**, D136–D143.



Fierer N., Hamady M., Lauber C.L., Knight R. (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 17994–17999.

Gilliam M. (1990) *Chalkbrood disease of honey bees, Apis mellifera, caused by the fungus, Ascosphaera apis: a review of past and current research*. Vth International Colloquium on Invertebrate Pathology and Microbial Control, Adelaide, Australia, pp 398–402.

Girard M., Chagnon M., Fournier V. (2012) Pollen diversity collected by honey bees in the vicinity of *Vaccinium* spp. crops and its importance for colony development This article is part of a Special Issue entitled 'Pollination biology research in Canada: perspectives on a mutualism at different scales'. *Botany*, **90**, 545–555.

Harpke D., Peterson A. (2008) 5.8 S motifs for the identification of pseudogenic ITS regions. *Botany*, **86**, 300–305.

Hebert P.D., Cywinska A., Ball S.L. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 313–321.

Hemmer W. (1997) *Foods derived from genetically modified organisms and detection methods*. Agency for Biosafety Research and Assessment of Technology Impacts of the Swiss Priority Programme Biotechnology of the Swiss National Science Foundation, http://www.bats.ch/bats/publikationen/1997-2_gmo/gmo_food.pdf.

Hollingsworth P.M., Graham S.W., Little D.P. (2011) Choosing and using a plant DNA barcode. *PLoS ONE*, **6**, e19254.

Keller A., Schleicher T., Schultz J., Müller T., Dandekar T., Wolf M. (2009) 5.8S–28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, **430**, 50–57.

Keller A., Grimmer G., Steffan-Dewenter I. (2013) Diverse microbiota identified in whole intact nest chambers of the red mason bee *Osmia bicornis* (Linnaeus 1758). *PLoS One*, **7829**, 6.

Kleijn D., Raemakers I. (2008) A retrospective analysis of pollen host plant use by stable and declining bumble bee species. *Ecology*, **89**, 1811–1823.

Koetschan C., Forster F., Keller A., Schleicher T., Ruderisch B., Schwarz R., Muller T., Wolf M., Schultz J.. (2010) The ITS2 Database III – sequences and structures for phylogeny. *Nucleic Acids Research*, **38**, D275–D279.

Kozich J.J., Westcott S.L., Baxter N.T., Highlander S.K., Schloss P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*, **79**, 5112–5120.

Kuczynski J., Stombaugh J., Walters W.A., González A., Caporaso J.G., Knight R. (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Current Protocols in Bioinformatics*, **10**, 17.

Kunin V., Engelbrektson A., Ochman H., Hugenholtz P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.

Lan Y., Wang Q., Cole J.R., Rosen G.L. (2012) Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS One*, **7**, e32491.

Liu K.-L., Porras-Alfaro A., Kuske C.R., Eichorst S.A., Xie G. (2012) Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Applied and Environmental Microbiology*, **78**, 1523–1533.

Marchant R., Behling H., Berrio J.C., Cleef A., Duivenvoorden J., Hooghiemstra H., Kuhry P., Melief B., Geel B.V., Van der Hammen T. (2001) Mid- to Late-Holocene pollen-based biome reconstructions for Colombia. *Quaternary Science Reviews*, **20**, 1289–1308.

Metzker M.L. (2009) Sequencing technologies—the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Mullins J., Emberlin J. (1997) Sampling pollens. *Journal of Aerosol Science*, **28**, 365–370.

Odoux J.-F., Feuillet D., Aupinel P., Loublier Y., Tasei J.-N., Mateescu C. (2012) Territorial biodiversity and consequences on physico-chemical characteristics of pollen collected by honey bee colonies. *Apidologie*, **43**, 561–575.

Parameswaran P., Jalili R., Tao L., Shokralla S., Gharizadeh B., Ronaghi M., Fire A.Z. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**, e130.

Picard-Nizou A., Pham-Delegue M., Kerguelen V., Douault P., Marilleau R., Olsen L., Grison R., Toppan A., Masson C. (1995) Foraging behaviour of honey bees (*Apis mellifera* L.) on transgenic oilseed rape (*Brassica napus* L. var. *oleifera*). *Transgenic Research*, **4**, 270–276.

R Development Core Team (2010) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (01/19).

Roberts R.J., Carneiro M.O., Schatz M.C. (2013) The advantages of SMRT sequencing. *Genome Biology*, **14**, 405.

Ruoff K., Luginbühl W., Kilchenmann V., Bosset J.O., von der Ohe K., von der Ohe W., Amadò R. (2007) Authentication of the botanical origin of honey using profiles of classical measurands and discriminant analysis. *Apidologie*, **38**, 438–452.

Schlumbaum A., Tensen M., Jaenicke-Després V. (2008) Ancient plant DNA in archaeobotany. *Vegetation History and Archaeobotany*, **17**, 233–244.

Schnell I.B., Fraser M., Willerslev E., Gilbert M.T.P. (2010) Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey. *Arthropod-Plant Interactions*, **4**, 107–116.

Sheffield C.S., Hebert P.D., Kevan P.G., Packer L. (2009) DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. *Molecular Ecology Resources*, **9** (Suppl. 1), 196–207.

Song J., Shi L., Li D., Sun Y., Niu Y., Chen Z., Luo H., Pang X., Sun Z., Liu C. (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. *PLoS One*, **7**, e43971.

Sowunmi M. (1976) The potential value of honey in palaeopalynology and archaeology. *Review of Palaeobotany and Palynology*, **21**, 171–185.

Spooner D.M. (2009) DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. *American Journal of Botany*, **96**, 1177–1189.

Staatliche Naturwissenschaftliche Sammlungen Bayern. (2013) Botanischer Informationsknoten Bayern, Germany.

Stillman E., Flenley J.R. (1996) The needs and prospects for automation in palynology. *Quaternary Science Reviews*, **15**, 1–5.

Sugita S. (1994) Pollen representation of vegetation in Quaternary sediments: theory and method in patchy vegetation. *Journal of Ecology*, **82**, 881–897.

Suzuki M.T., Giovannoni S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.

Taylor H.R., Harris W.E. (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, **12**, 377–388.

Tzedakis P. (1993) Long-term tree populations in northwest Greece through multiple Quaternary climatic cycles. *Nature*, **364**, 437–440.

Valentini A., Pompanon F., Taberlet P. (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.

Valentini A., Miquel C., Taberlet P. (2010) DNA barcoding for honey biodiversity. *Diversity*, **2**, 610–617.

Velasco R., Zharkikh A., Troggio M., Cartwright D.A., Cestaro A., Pruss D., Pindo M., FitzGerald L.M., Vezzulli S., Reid J. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**, e1326.

Wang Q., Garrity G.M., Tiedje J.M., Cole J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Wcislo W.T., Cane J.H. (1996) Floral resource utilization by solitary bees (Hymenoptera: Apoidea) and exploitation of their stored foods by natural enemies. *Annual Review in Entomology*, **41**, 257–286.

White T., Bruns T., Lee S., Taylor J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis M., Gelfand D., Shinsky J., White T. (Eds), *PCR-protocols: a Guide to Methods and Applications*. Academic Press, San Diego, pp 315–322.

Wilcock C., Neiland R. (2002) Pollination failure in plants: why it happens and when it matters. *Trends in Plant Science*, **7**, 270–277.

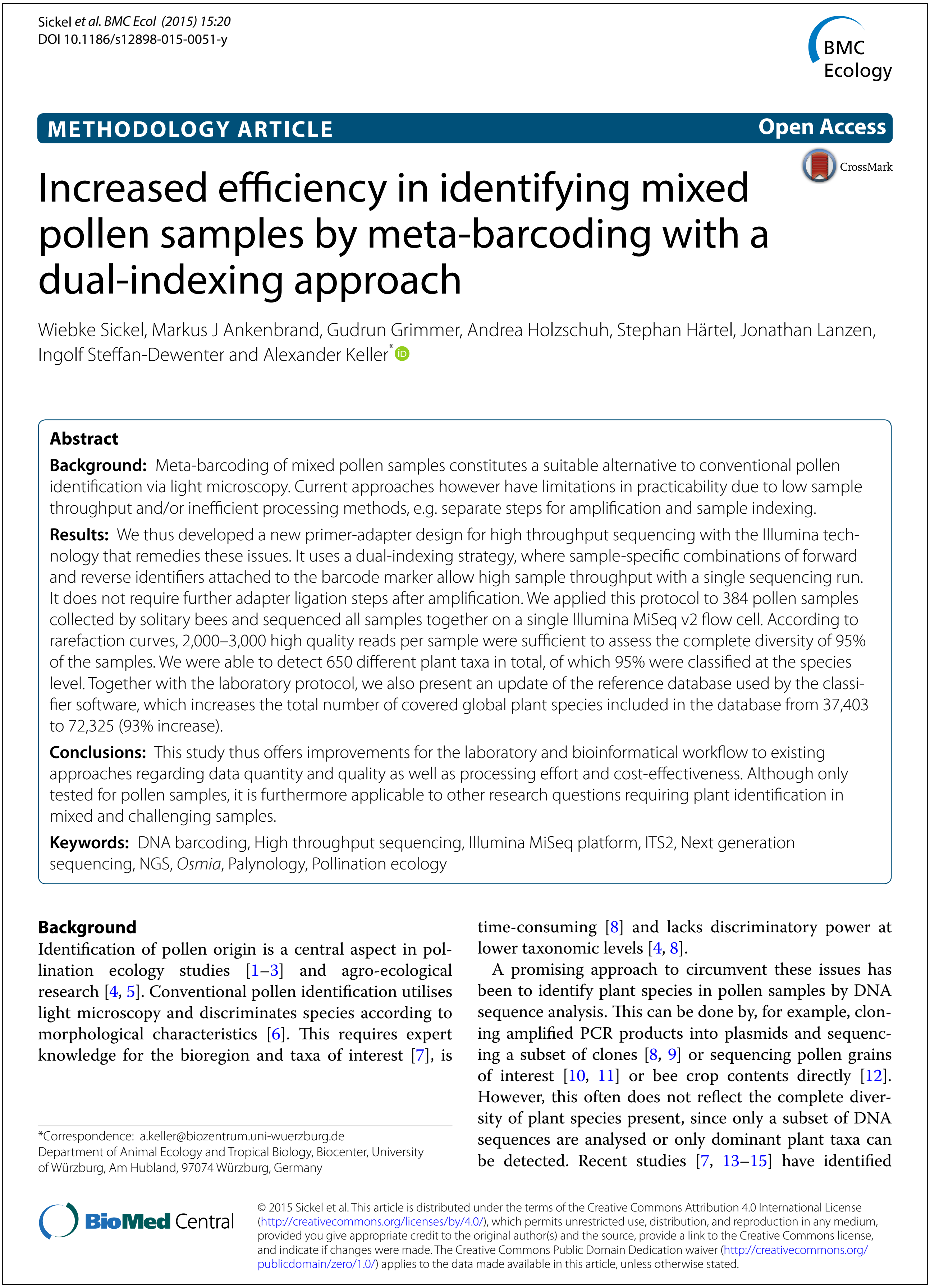
Wilson E.E., Sidhu C.S., LeVan K.E., Holway D.A. (2010) Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. *Molecular Ecology*, **19**, 4823–4829.

Zhou L.J., Pei K.Q., Zhou B., Ma K.P. (2007) A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil. *American Journal of Botany*, **94**, 477–481.

5.3 INCREASED EFFICIENCY IN IDENTIFYING MIXED POLLEN SAMPLES BY META-BARCODING WITH A DUAL-INDEXING APPROACH

– published in *BMC Ecology* –

This publication is also part of Sickel (2017). It is part of both theses because of complementary contributions.



Sickel *et al. BMC Ecol (2015) 15:20*
DOI 10.1186/s12898-015-0051-y

BMC Ecology

METHODOLOGY ARTICLE Open Access

Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach

Wiebke Sickel, Markus J Ankenbrand, Gudrun Grimmer, Andrea Holzschuh, Stephan Härtel, Jonathan Lanzen, Ingolf Steffan-Dewenter and Alexander Keller*

Abstract

Background: Meta-barcoding of mixed pollen samples constitutes a suitable alternative to conventional pollen identification via light microscopy. Current approaches however have limitations in practicability due to low sample throughput and/or inefficient processing methods, e.g. separate steps for amplification and sample indexing.

Results: We thus developed a new primer-adapter design for high throughput sequencing with the Illumina technology that remedies these issues. It uses a dual-indexing strategy, where sample-specific combinations of forward and reverse identifiers attached to the barcode marker allow high sample throughput with a single sequencing run. It does not require further adapter ligation steps after amplification. We applied this protocol to 384 pollen samples collected by solitary bees and sequenced all samples together on a single Illumina MiSeq v2 flow cell. According to rarefaction curves, 2,000–3,000 high quality reads per sample were sufficient to assess the complete diversity of 95% of the samples. We were able to detect 650 different plant taxa in total, of which 95% were classified at the species level. Together with the laboratory protocol, we also present an update of the reference database used by the classifier software, which increases the total number of covered global plant species included in the database from 37,403 to 72,325 (93% increase).

Conclusions: This study thus offers improvements for the laboratory and bioinformatical workflow to existing approaches regarding data quantity and quality as well as processing effort and cost-effectiveness. Although only tested for pollen samples, it is furthermore applicable to other research questions requiring plant identification in mixed and challenging samples.

Keywords: DNA barcoding, High throughput sequencing, Illumina MiSeq platform, ITS2, Next generation sequencing, NGS, *Osmia*, Palynology, Pollination ecology

Background

Identification of pollen origin is a central aspect in pollination ecology studies [1–3] and agro-ecological research [4, 5]. Conventional pollen identification utilises light microscopy and discriminates species according to morphological characteristics [6]. This requires expert knowledge for the bioregion and taxa of interest [7], is time-consuming [8] and lacks discriminatory power at lower taxonomic levels [4, 8].

A promising approach to circumvent these issues has been to identify plant species in pollen samples by DNA sequence analysis. This can be done by, for example, cloning amplified PCR products into plasmids and sequencing a subset of clones [8, 9] or sequencing pollen grains of interest [10, 11] or bee crop contents directly [12]. However, this often does not reflect the complete diversity of plant species present, since only a subset of DNA sequences are analysed or only dominant plant taxa can be detected. Recent studies [7, 13–15] have identified

*Correspondence: a.keller@biozentrum.uni-wuerzburg.de
Department of Animal Ecology and Tropical Biology, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

high throughput sequencing (HTS) approaches based on meta-barcoding as a suitable alternative for existing methods. However, current protocols still suffer from a limited sample throughput [7, 14, 15] and/or practicability issues due to separate steps for PCR amplification and index labelling [13]. We here present a protocol for highly multiplexed pollen sequencing utilising a dual-indexing strategy [16]. An overview of existing methods alongside our new approach is given in Figure 1. We designed meta-barcoding primers suitable for plant identification using the internal transcribed spacer 2 (ITS2) that already incorporate Illumina-specific adapters for high-throughput sequencing as well as new sequencing primers that are added to the sequencing flow cell. The rationale for using ITS2 rather than other genetic markers for plant DNA barcoding in general is provided elsewhere [17] and its applicability regarding meta-barcoding criteria has also been successfully demonstrated [7, 13]. We tested our new approach by sequencing 384 pollen samples collected by two solitary bees species with known different foraging strategies: polylectic *Osmia bicornis* [18] and oligolectic *Osmia truncorum* [19]. Alongside this enhancement of the laboratory method, we updated the reference database used for ITS2 meta-barcoding [7] and added compatibility for the UTAX classification software [20] as a second and alternative strategy beside the RDP classifier [7, 21].

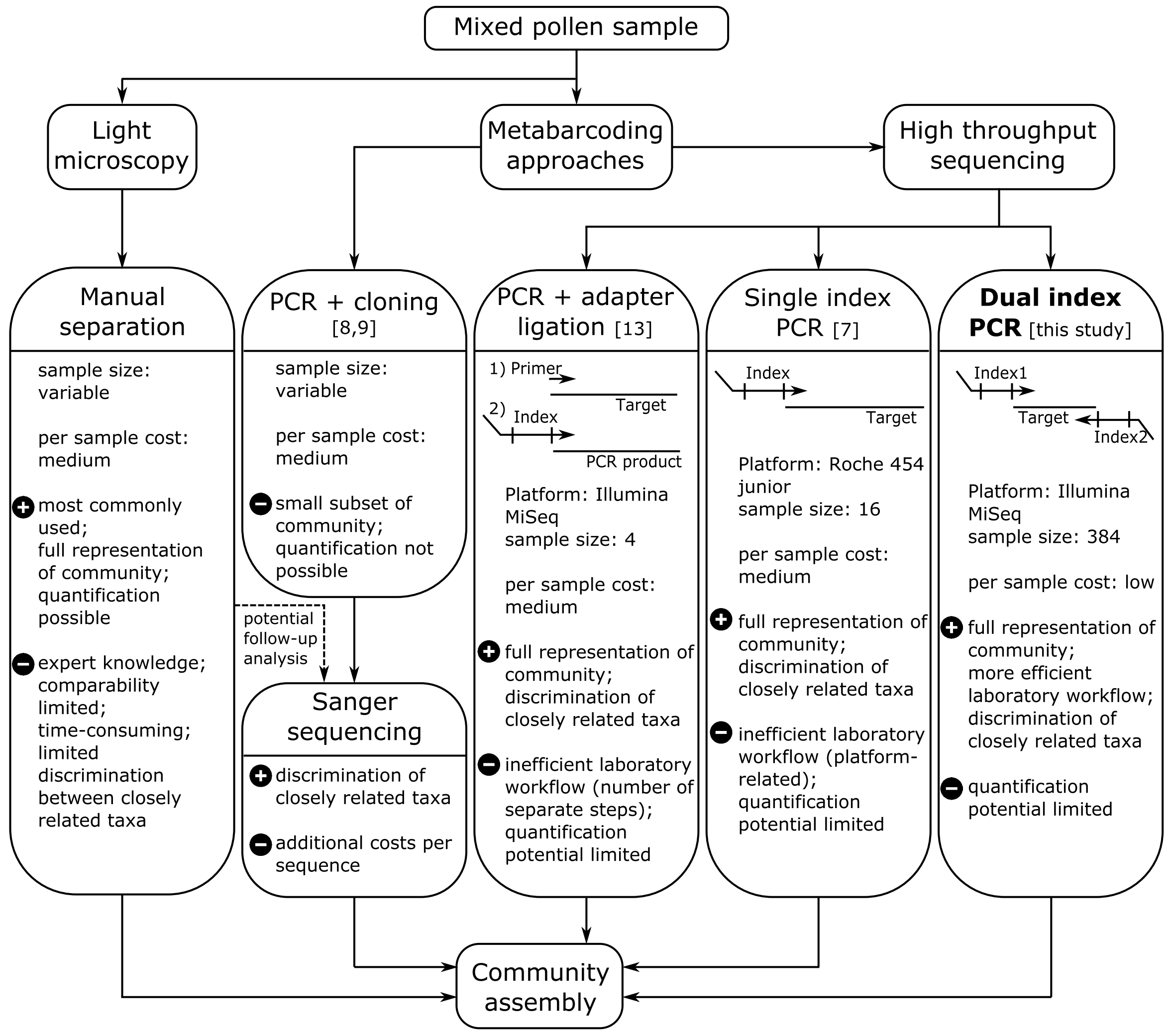
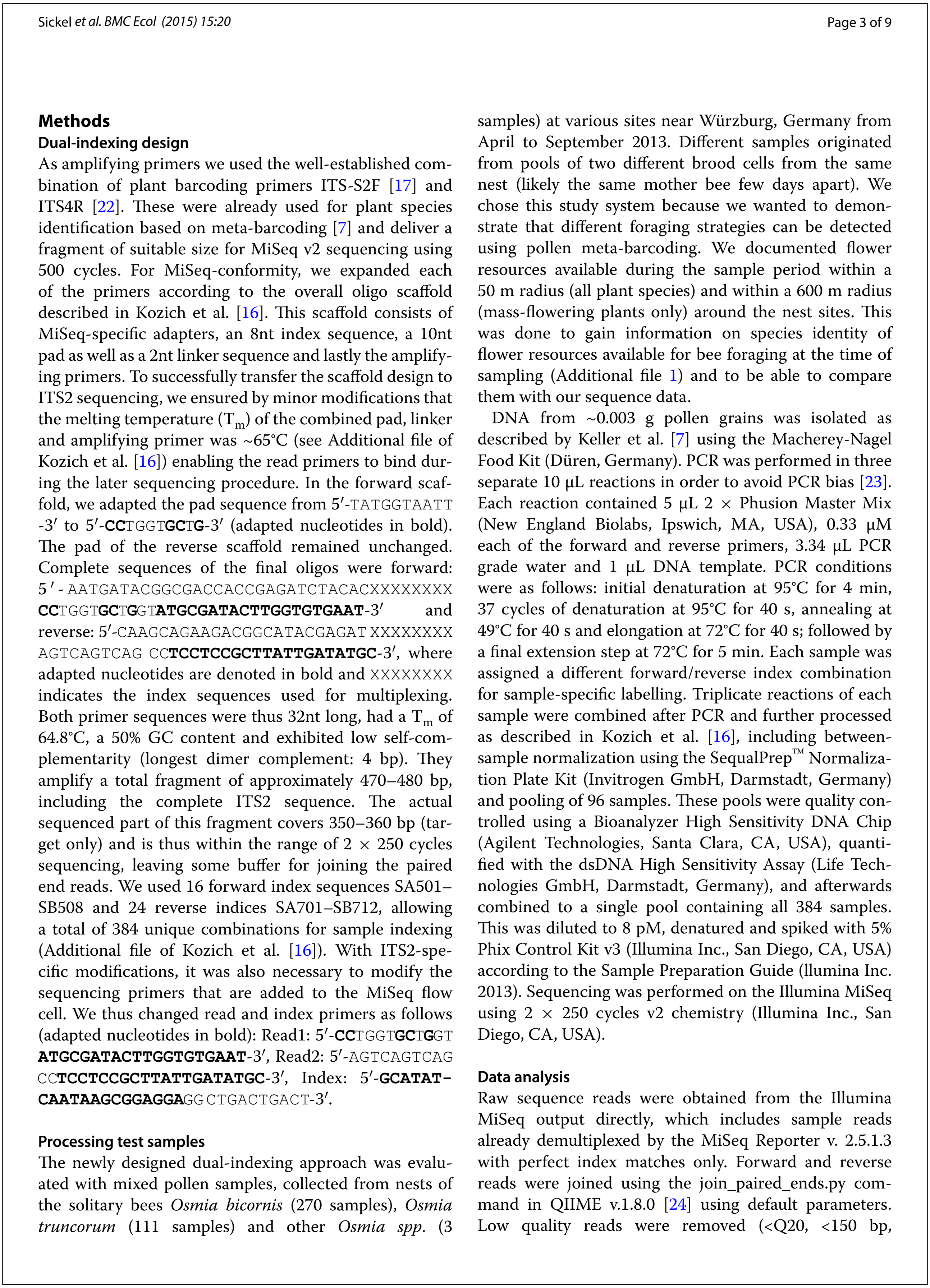

Mixed pollen sample
Light microscopy
Metabarcoding approaches
High throughput sequencing
Manual separation
sample size: variable
per sample cost: medium
+ most commonly used; full representation of community; quantification possible
− expert knowledge; comparability limited; time-consuming; limited discrimination between closely related taxa
PCR + cloning [8,9]
sample size: variable
per sample cost: medium
− small subset of community; quantification not possible
potential follow-up analysis
Sanger sequencing
+ discrimination of closely related taxa
− additional costs per sequence
PCR + adapter ligation [13]
1) Primer
Target
2) Index
PCR product
Platform: Illumina MiSeq
sample size: 4
per sample cost: medium
+ full representation of community; discrimination of closely related taxa
− inefficient laboratory workflow (number of separate steps); quantification potential limited
Single index PCR [7]
Index
Target
Platform: Roche 454 junior
sample size: 16
per sample cost: medium
+ full representation of community; discrimination of closely related taxa
− inefficient laboratory workflow (platform-related); quantification potential limited
Dual index PCR [this study]
Index1
Target
Index2
Platform: Illumina MiSeq
sample size: 384
per sample cost: low
+ full representation of community; more efficient laboratory workflow; discrimination of closely related taxa
− quantification potential limited
Community assembly

Figure 1 Comparison of different approaches for plant species identification in mixed pollen samples.



Methods

Dual-indexing design

As amplifying primers we used the well-established combination of plant barcoding primers ITS-S2F [17] and ITS4R [22]. These were already used for plant species identification based on meta-barcoding [7] and deliver a fragment of suitable size for MiSeq v2 sequencing using 500 cycles. For MiSeq-conformity, we expanded each of the primers according to the overall oligo scaffold described in Kozich et al. [16]. This scaffold consists of MiSeq-specific adapters, an 8nt index sequence, a 10nt pad as well as a 2nt linker sequence and lastly the amplifying primers. To successfully transfer the scaffold design to ITS2 sequencing, we ensured by minor modifications that the melting temperature (T_m) of the combined pad, linker and amplifying primer was ~65°C (see Additional file of Kozich et al. [16]) enabling the read primers to bind during the later sequencing procedure. In the forward scaffold, we adapted the pad sequence from 5′-TATGGTAATT -3′ to 5′-**CC**TGGT**GC**T**G**-3′ (adapted nucleotides in bold). The pad of the reverse scaffold remained unchanged. Complete sequences of the final oligos were forward: 5′-AATGATACGGCGACCACCGAGATCTACACXXXXXXXX **CC**TGGT**GC**T**G**GT**ATGCGATACTTGGTGTGAAT**-3′ and reverse: 5′-CAAGCAGAAGACGGCATACGAGAT XXXXXXXX AGTCAGTCAG CC**TCCTCCGCTTATTGATATGC**-3′, where adapted nucleotides are denoted in bold and XXXXXXXX indicates the index sequences used for multiplexing. Both primer sequences were thus 32nt long, had a T_m of 64.8°C, a 50% GC content and exhibited low self-complementarity (longest dimer complement: 4 bp). They amplify a total fragment of approximately 470–480 bp, including the complete ITS2 sequence. The actual sequenced part of this fragment covers 350–360 bp (target only) and is thus within the range of 2 × 250 cycles sequencing, leaving some buffer for joining the paired end reads. We used 16 forward index sequences SA501–SB508 and 24 reverse indices SA701–SB712, allowing a total of 384 unique combinations for sample indexing (Additional file of Kozich et al. [16]). With ITS2-specific modifications, it was also necessary to modify the sequencing primers that are added to the MiSeq flow cell. We thus changed read and index primers as follows (adapted nucleotides in bold): Read1: 5′-**CC**TGGT**GC**T**G**GT **ATGCGATACTTGGTGTGAAT**-3′, Read2: 5′-AGTCAGTCAG CC**TCCTCCGCTTATTGATATGC**-3′, Index: 5′-**GCATATCAATAAGCGGAGGA**GG CTGACTGACT-3′.

Processing test samples

The newly designed dual-indexing approach was evaluated with mixed pollen samples, collected from nests of the solitary bees *Osmia bicornis* (270 samples), *Osmia truncorum* (111 samples) and other *Osmia spp.* (3 samples) at various sites near Würzburg, Germany from April to September 2013. Different samples originated from pools of two different brood cells from the same nest (likely the same mother bee few days apart). We chose this study system because we wanted to demonstrate that different foraging strategies can be detected using pollen meta-barcoding. We documented flower resources available during the sample period within a 50 m radius (all plant species) and within a 600 m radius (mass-flowering plants only) around the nest sites. This was done to gain information on species identity of flower resources available for bee foraging at the time of sampling (Additional file 1) and to be able to compare them with our sequence data.

DNA from ~0.003 g pollen grains was isolated as described by Keller et al. [7] using the Macherey-Nagel Food Kit (Düren, Germany). PCR was performed in three separate 10 μL reactions in order to avoid PCR bias [23]. Each reaction contained 5 μL 2 × Phusion Master Mix (New England Biolabs, Ipswich, MA, USA), 0.33 μM each of the forward and reverse primers, 3.34 μL PCR grade water and 1 μL DNA template. PCR conditions were as follows: initial denaturation at 95°C for 4 min, 37 cycles of denaturation at 95°C for 40 s, annealing at 49°C for 40 s and elongation at 72°C for 40 s; followed by a final extension step at 72°C for 5 min. Each sample was assigned a different forward/reverse index combination for sample-specific labelling. Triplicate reactions of each sample were combined after PCR and further processed as described in Kozich et al. [16], including between-sample normalization using the SequalPrep™ Normalization Plate Kit (Invitrogen GmbH, Darmstadt, Germany) and pooling of 96 samples. These pools were quality controlled using a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA, USA), quantified with the dsDNA High Sensitivity Assay (Life Technologies GmbH, Darmstadt, Germany), and afterwards combined to a single pool containing all 384 samples. This was diluted to 8 pM, denatured and spiked with 5% Phix Control Kit v3 (Illumina Inc., San Diego, CA, USA) according to the Sample Preparation Guide (llumina Inc. 2013). Sequencing was performed on the Illumina MiSeq using 2 × 250 cycles v2 chemistry (Illumina Inc., San Diego, CA, USA).

Data analysis

Raw sequence reads were obtained from the Illumina MiSeq output directly, which includes sample reads already demultiplexed by the MiSeq Reporter v. 2.5.1.3 with perfect index matches only. Forward and reverse reads were joined using the join_paired_ends.py command in QIIME v.1.8.0 [24] using default parameters. Low quality reads were removed (<Q20, <150 bp, ambiguous base-pairs) with USEARCH v8.0.1477 [25]. Combined reads were taxonomically classified with the RDP classifier [21] as well as the UTAX algorithm and results compared to show that the data is compatible between both alternative analytical strategies. UTAX and RDP were executed for each sample separately.

In the following, we concentrate on UTAX, since the RDP classifier has been used previously for pollen taxonomic assignments [7]. A raw score cut-off at 20 was used, as the UTAX algorithm does currently not provide bootstrap comparable confidence values (but is expected to incorporate these soon, see http://drive5.com/usearch/manual/faq_taxconfs.html, accessed 2015/22/05). These assignment scores are however comparable between reads as long as subsequent analyses do base all upon the same database.

For data analysis, the raw UTAX output was parsed using a self-written perl script, which counts the number of assignments for each taxon and aggregates these into a single table (https://github.com/iimog/meta-barcoding-dual-indexing). This table is converted into a community matrix format, with rows as species and columns representing samples, and a separate file with the taxonomic lineage of each species is also created. These files are directly importable into common statistical software, e.g. *R* v.3.1.2 [26] using the package *phyloseq* v.1.6.1 [27]. To assess sufficiency of the sequencing depth, we created species accumulation curves for each sample using the *vegan* package v2.2-0 [28] in *R* v.3.1.2 [26], excluding taxa accounting for less than 0.1% of sample reads. Additionally, we determined the ten most abundant plant families collected by *O. bicornis* and *O. truncorum*.

Reference database update

Beside the enhancement of the laboratory protocol, we considered it important to address also the actuality and completeness of the reference database. We thus performed an update according to the annotation pipeline described for the ITS2 database [29, 30]. For this, we extracted all available ITS2 sequences belonging to Viridiplantae from GenBank [31] (accessed on 2015/19/01) as described in detail in Koetschan et al. [30]. The taxonomy follows the NCBI taxonomy database [32], which may not perfectly reflect evolutionary status, but is well usable for automatic procedures, due to its integration into the public NCBI framework. Taxonomy was assigned to the sequences by mapping the gi to the NCBI taxid. Taxonomic levels were selected at seven levels (kingdom, phylum, class, order, family, genus, species) using a custom perl script utilizing the NCBI::Taxonomy module by courtesy of F. Förster (doi:10.5281/zenodo.17375). RDP training files, a UTAX database and taxtree were created with a custom perl script (https://github.com/iimog/meta-barcoding-dual-indexing). The database update, scripts and information on how to use it with the RDP classifier or UTAX are provided at http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

Results

Sequencing output and data analysis

In total we obtained 11,624,087 raw ITS2 reads (PhiX excluded), which accounted for an average of 30,271 [standard deviation (SD): 11,373; median: 30,900] reads per sample. After data processing (low-quality <Q20, short reads <150 bp, ambiguous base-pairs), a mean of 15,580 (SD 6,598; median 15,740) reads per sample remained. Species accumulation curves (Figure 2) show that almost all samples were sequenced to saturation after approximately 2,000–3,000 high quality reads. Based on the ratio of raw to high quality reads, this accounts for approximately 4,000–6,000 raw reads required. Per sample pollen in bee brood cells originated from between one and 85 different plant species (Figure 2). Five per cent of samples (19) yielded an output of less than 2,000 reads (minimum saturation threshold, Figure 2), which were removed prior to further analysis. Raw sequences are accessible via the EBI-SRA with the project accession number PRJEB8640.

Reference database update

Our previously published database contained 73,853 reference sequences of 37,403 unique plant species [7]. The updated version now contains 182,505 plant sequences from 72,325 different species. This is an increase by factor 2.47 (147% additional) for sequences and 1.93 (93% additional) for unique species. In comparison with the original reference set [7], with these data 80.1% (original 53.1%) of the plant species and 90.4% (original 75%) of the genera in Bavaria, Germany, where our test samples originate from, were covered (data retrieved from http://bayernflora.de; accessed on 2015/01/24). Correspondingly, for plant species in the USA, the database covers 66.5–79.1% (median 76.1%) of species and 73.8–87.3% (median 84.9%) of genera, depending on the US state (data retrieved from the BISON project; http://bison.usgs.ornl.gov; accessed on 2015/04/02). In both cases, Bavaria and USA, missing species are likely rare or endemic to specific regions. A comparison of numbers of genera per order covered in the old and updated database versions can be found in the Additional file 2: Table S1.

Test samples

Regarding our samples, taxonomic classification (after filtering out rare taxa below 0.1%) identified 650 different plant taxa, of which 617 could be classified taxonomically to plant species level, belonging to 288 genera, 71

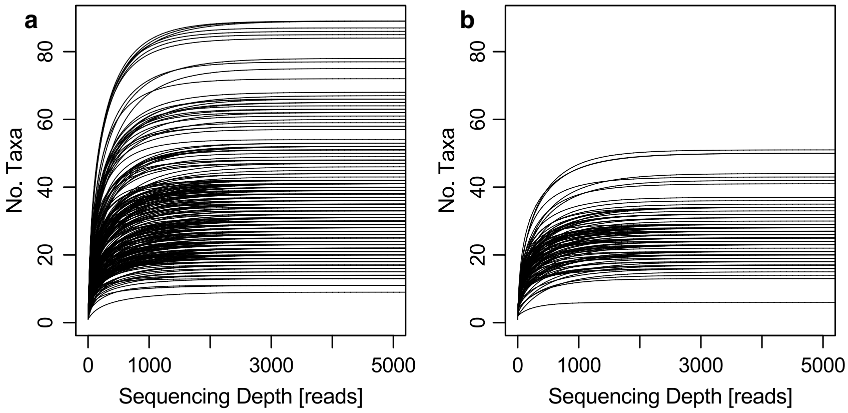


Figure 2 Species accumulation curves. **a** *Osmia bicornis* samples; **b** *Osmia truncorum* samples. The x-axis was limited to 5,000 reads as the saturation of all samples was below this threshold. The y-axis was limited to 90 taxa in both plots to obtain the same scale. Taxa accounting for less than 0.1% of total sample reads were excluded.

families, 37 orders and nine classes. The remaining 33 taxa (5%) could not be classified at the species level. Of these, 17 taxa could still be classified at genus level and another seven at the family level. Nine taxa remained that could not be classified even to family level. These belonged to the Sapindales, Fagales and Microthamniales (one taxon each) or remained unclassified (six taxa). At the genus level, RDP and UTAX taxonomic assignments agreed in ~90% of all read classifications, thus both classifiers yielded comparable results.

For both *Osmia* species together, approximately 50% of documented plant genera (<50 m: all plants, 50–600 m: only mass-flowering plants) were detectable within the sequencing data and contributed with ~75% to all quality-filtered reads. The two bee species differed clearly in foraging patterns as visible through plant families predominantly collected (Figure 3), as well as in the number of plant species with *O. bicornis* collecting up to 85 plant species and *O. truncorum* collecting up to 50 plant species per brood cell (Figure 2). The ten most abundant plant families collected by *O. bicornis* were Brassicaceae (27.07%), Ranunculaceae (16.98%), Aceraceae (11.62%), Fagaceae (10.86%), Juglandaceae (7.16%), Papaveraceae (5.91%) Fabaceae (5.40%), Asteraceae (4.89%), Rosaceae (3.59%) and Plantaginaceae (2.62%). *O. truncorum* pollen was dominated by Asteraceae (92.92%), and only Caprifoliaceae (1.51%) and Brassicaceae (1.14%) contributed more than 1% to the overall collection. The Asteraceae collected by *O. truncorum* contained a wide spectrum of plant genera, with 58 genera being detected, the ten most abundant of which were *Picris, Jacobaea, Tanacetum, Artemisia, Achillea, Tripleurospermum, Inula, Cota, Leucanthemum* and *Crepis* (Figure 3).

Discussion

High throughput sequencing (HTS) has been shown to be successful and valuable for taxonomic assessment of mixed pollen samples [7, 13, 15]. The drawbacks of existing protocols were the low number of samples processed simultaneously or inefficient multistep library preparations. Recent developments in sequencing technologies allow far larger multiplexing, given the enormous throughput already available with desktop NGS devices. Highly multiplexed sample processing has already been established for bacterial assessments using dual-indexing approaches with the MiSeq sequencer [16]. It was the goal of this study to transfer this knowledge to the field of plant meta-barcoding, in our specific case of pollen samples.

By adapting the primer design to the ITS2 region, modifying the oligo scaffold design, and adjusting the sequencing primers to be compatible with the MiSeq device, we successfully established a fast pollen DNA meta-barcoding routine with high multiplexing capabilities. For our test samples, the newly designed primers were used to sequence 384 mixed pollen samples collected by solitary bees with a single sequencing run. In the original bacterial dual-indexing protocol [16], the potential for higher multiplex rates than 384 samples is suggested depending on required throughput to assess the diversity. Our sequencing results indicate that for pollen samples at least a depth of 2,000–3,000 high quality reads per sample should be reached to identify all taxa within the sample (plateau reached, Figure 2), which was comparable for the two bee species under study. However, this is of course highly dependent on number of plant species in the samples, which may be dependent on

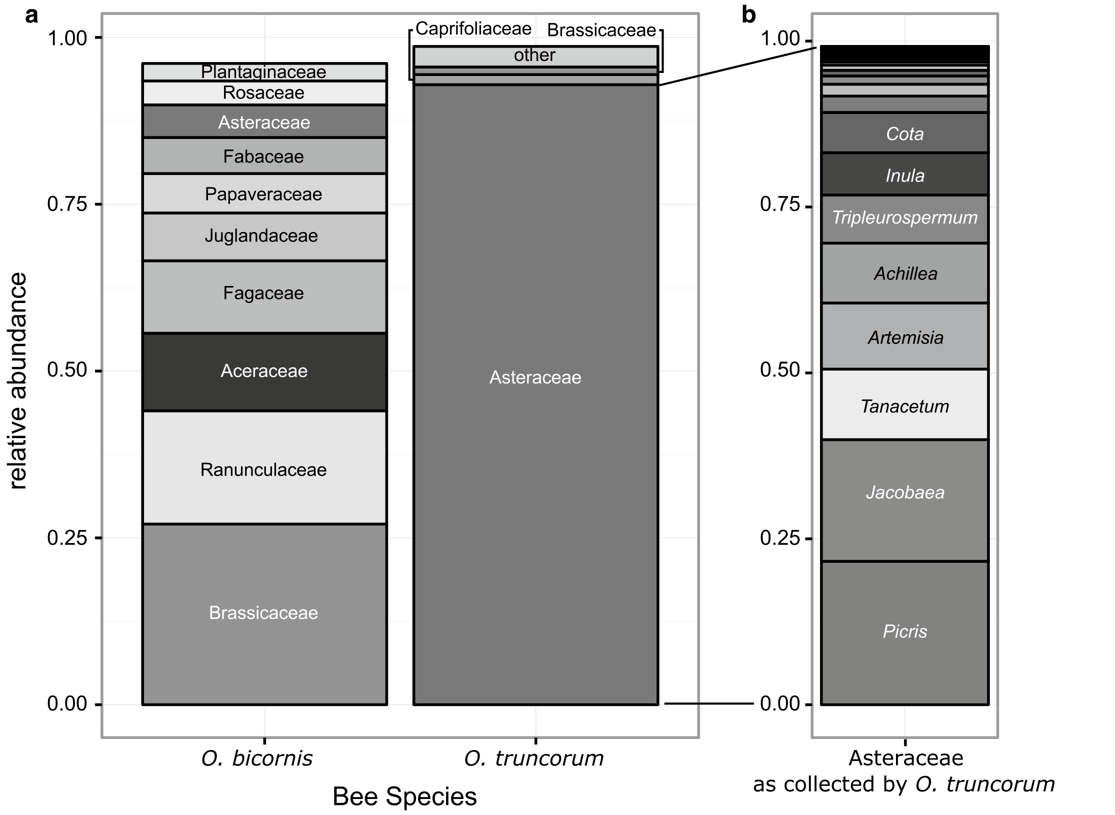


Figure 3 Pollen spectrum of the two bee species. **a** Ten most abundant families as collected by the bee species *O. bicornis* and *O. truncorum*. For *O. truncorum* 'other' include the families Apiaceae, Rosaceae, Fabaceae, Ranunculaceae, Plantaginaceae, Juglandaceae and Amaranthaceae. **b** Plant genera detected within the Asteraceae collected by *O. truncorum*. For visualisation reasons, only the eight most abundant genera are labelled. Please note that Aceraceae is now included within Sapindaceae.

sample origin, foraging behaviour and the biodiversity of the ecosystem of interest, but may serve nonetheless as a guideline for higher multiplex rates. Additional index combinations for more samples are provided in the Additional files alongside the protocol for the bacterial dual-index approach [16].

Beside our dual-indexing strategy, another HTS-based approach has been recently proposed. There, PCR amplification and index labelling were conducted in separate steps [13], which is time and labour-intensive and introduces a further step where errors may be introduced. In our protocol, PCR amplification and sample indexing occur simultaneously, which is highly practical and requires no special reagents, such as additional expensive library preparation kits or adapter ligation chemicals. In our protocol, the complete workflow accounts for less than USD 20.00 for materials per sample, when processing 384 samples simultaneously. This is much lower than conventional pollen analysis under the light microscope, which can reach several hundred USD per sample.

Most plant taxa detected could be successfully classified using the already shown RDP classifier [7, 21], but also the recently developed UTAX algorithm [25]. Due to the missing confidence values for taxonomic assignments in UTAX version 8.0 (announced for version 8.1, http://drive5.com/usearch/manual/faq_taxconfs.html, accessed 2015/22/05), we compared the classifications to the RDP output as well as the documented flower resources. UTAX and RDP showed high agreement between taxonomic classifications, thus both may be used arbitrarily.

Approximately half of the genera found flowering near the nest sites were detected in the pollen samples. This is attributable to bee foraging preferences, where not all available resources might be used, especially for the oligolectic *O. truncorum*. Secondly, about three quarters of the reads were assigned to plant genera documented near the nesting sites (<50 m: all plant species, 50–600 m:

mass-flowering plants only). As bees are expected to forage also further away, the remaining reads are attributable to pollen collected from undocumented plants or misclassifications.

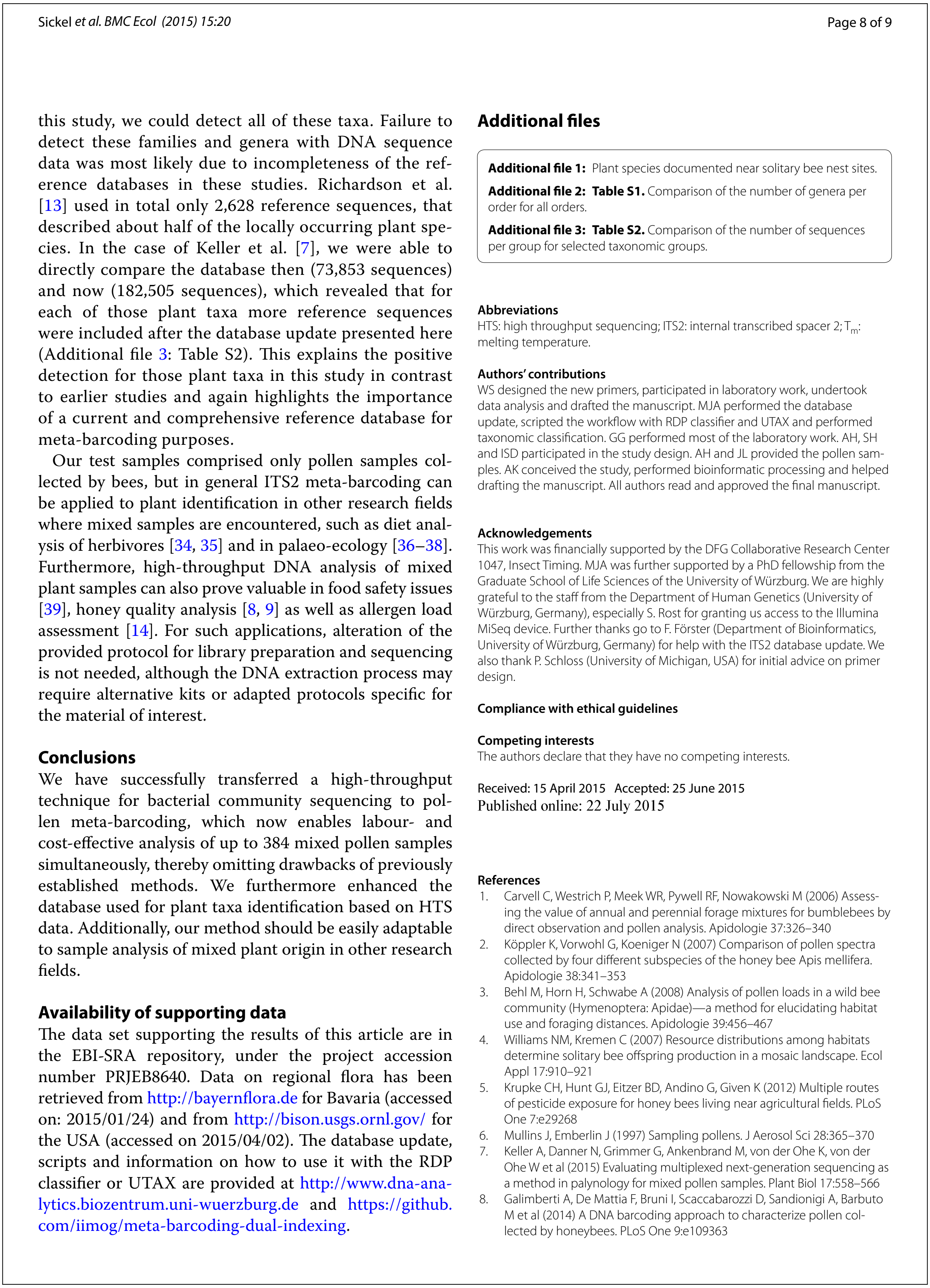
According to our expectation, pollen composition patterns were very different for the oligolectic and the polylectic bee species (Figure 3). *O. truncorum* samples were dominated by Asteraceae, whereas *O. bicornis* samples showed a wide pollen spectrum. Our data correspond to flower preferences and foraging strategies known for these species [18, 19]. This supports the high quality of information obtained by pollen meta-barcoding, as already intensively evaluated in another study [7]. It is noteworthy that even very rare taxa could be detected, which is of special interest in the oligolectic *O. truncorum* and might be overlooked in light microscopy assessment of pollen samples.

We would like to point out that abundance data obtained from molecular approaches should in general be interpreted with care and only as relative abundance (divided by total number of reads in the sample to account for varying library sizes). Contradicting results exist concerning the suitability of pollen meta-barcoding for quantification purposes, with Keller et al. [7] and Kraaijeveld et al. [14] finding a positive significant correlation between genera by light microscopy and meta-barcoding, whilst Richardson et al. [13] were not able to find such a connection. Due to the different steps in the workflow, e.g. dilutions and PCR, biases can be introduced, leading to skewed data and over- or underrepresentation of certain taxa. PCR bias is considered to be a random process and can be accounted for by performing replicate PCR reactions for each sample [23], which are pooled subsequently. We followed this approach in this study likewise to Keller et al. [7] to avoid PCR bias as far as possible. This may explain some of the discrepancy between studies, although a recent study indicated that PCR replicates might not be necessary in pollen meta-barcoding [14]. The reduced amount of individual processing steps of direct indexing, (as performed here and in both studies identifying positive correlation [7, 14]) further reduces additional risks to introduce unwanted effects in comparison with the study using adapter ligation that shows no correlation [13].

In this study, samples of the same bee species show high consistency in abundance patterns of major taxa, which are easily biologically explainable. A good compromise for most studies investigating foraging patterns might be to not use direct count data, but conservatively categorising plant taxa into 'abundant' and 'rare' based on a threshold, as proposed by Keller et al. [7]. Where more detail is needed, a subset of samples may also be analysed in parallel by light microscopy for evaluation purposes [7, 13, 14].

One major advantage of pollen meta-barcoding is that no expert knowledge on pollen morphology is required for taxonomic assignment. Additionally, species level assignment is possible even for closely related plant taxa. However, successful taxonomic assignment critically depends on the quality of the reference database. Our target marker was the ITS2 region, but other genetic markers might also be considered for plant species identification using meta-barcoding, e.g. *trn*L [14, 15] or *rbcL* plus *trnH-psbA* [8, 9]. The described dual indexing approach [16] can also be applied to other genetic markers, provided some considerations are taken into account as described for ITS2 in this study. On the laboratory side of the workflow, firstly target and thereby primer choice should be appropriate for universal amplification and plant species identification based on DNA sequence data. The amplified fragment should be of the appropriate size for the chosen MiSeq sequencing chemistry, e.g. no longer than ~480–490 bp for 2 × 250 v2 sequencing kits, allowing for some overlap between forward and reverse reads. Given these conditions are met, primer design can be performed following the guidelines from Kozich et al. [16] including the required modifications to the various oligonucleotides. However, as mentioned before, successful plant species identification relies to a large degree also on the underlying reference database and bioinformatical classification algorithm. For most alternative markers comprehensive reference databases are currently lacking and thus taxonomic classifications are mainly performed by a BLAST search [33] against sequences downloaded from GenBank [8, 9, 13–15], locally managed alternative databases [9] and/or newly acquired DNA sequences [8, 9]. BLAST searches are based on local alignments that may only use parts of each sequence (e.g. conserved regions) for classification, lack a hierarchy classification procedure and results can be difficult to interpret [7, 17] especially when results show hits for multiple, different taxa. Setting up locally managed databases is time- and labour-intensive a well as costly and makes it difficult to compare independent studies with one another. In the case of the ITS2 region, we benefitted from the already established ITS2 database [30], which contains annotated and trimmed ITS2 sequences from species worldwide and can be publicly accessed, improving overall comparability across studies.

Although Chen et al. [17] reported high identification accuracies with ITS2 as a genetic marker, some plant taxa could not be identified in recent studies on pollen meta-barcoding [7, 13]. These included the families Salicaceae, Lamiaceae [13] and Vitaceae [7] and the genera *Lonicera* [13], *Heracleum*, *Carduus*, *Phacelia*, *Convolvulus* and *Helianthus* [7], although they had been identified with microscopic pollen analysis. In



this study, we could detect all of these taxa. Failure to detect these families and genera with DNA sequence data was most likely due to incompleteness of the reference databases in these studies. Richardson et al. [13] used in total only 2,628 reference sequences, that described about half of the locally occurring plant species. In the case of Keller et al. [7], we were able to directly compare the database then (73,853 sequences) and now (182,505 sequences), which revealed that for each of those plant taxa more reference sequences were included after the database update presented here (Additional file 3: Table S2). This explains the positive detection for those plant taxa in this study in contrast to earlier studies and again highlights the importance of a current and comprehensive reference database for meta-barcoding purposes.

Our test samples comprised only pollen samples collected by bees, but in general ITS2 meta-barcoding can be applied to plant identification in other research fields where mixed samples are encountered, such as diet analysis of herbivores [34, 35] and in palaeo-ecology [36–38]. Furthermore, high-throughput DNA analysis of mixed plant samples can also prove valuable in food safety issues [39], honey quality analysis [8, 9] as well as allergen load assessment [14]. For such applications, alteration of the provided protocol for library preparation and sequencing is not needed, although the DNA extraction process may require alternative kits or adapted protocols specific for the material of interest.

Conclusions

We have successfully transferred a high-throughput technique for bacterial community sequencing to pollen meta-barcoding, which now enables labour- and cost-effective analysis of up to 384 mixed pollen samples simultaneously, thereby omitting drawbacks of previously established methods. We furthermore enhanced the database used for plant taxa identification based on HTS data. Additionally, our method should be easily adaptable to sample analysis of mixed plant origin in other research fields.

Availability of supporting data

The data set supporting the results of this article are in the EBI-SRA repository, under the project accession number PRJEB8640. Data on regional flora has been retrieved from http://bayernflora.de for Bavaria (accessed on: 2015/01/24) and from http://bison.usgs.ornl.gov/ for the USA (accessed on 2015/04/02). The database update, scripts and information on how to use it with the RDP classifier or UTAX are provided at http://www.dna-analytics.biozentrum.uni-wuerzburg.de and https://github.com/iimog/meta-barcoding-dual-indexing.

Additional files

Additional file 1: Plant species documented near solitary bee nest sites.

Additional file 2: Table S1. Comparison of the number of genera per order for all orders.

Additional file 3: Table S2. Comparison of the number of sequences per group for selected taxonomic groups.

Abbreviations

HTS: high throughput sequencing; ITS2: internal transcribed spacer 2; T_m: melting temperature.

Authors' contributions

WS designed the new primers, participated in laboratory work, undertook data analysis and drafted the manuscript. MJA performed the database update, scripted the workflow with RDP classifier and UTAX and performed taxonomic classification. GG performed most of the laboratory work. AH, SH and ISD participated in the study design. AH and JL provided the pollen samples. AK conceived the study, performed bioinformatic processing and helped drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was financially supported by the DFG Collaborative Research Center 1047, Insect Timing. MJA was further supported by a PhD fellowship from the Graduate School of Life Sciences of the University of Würzburg. We are highly grateful to the staff from the Department of Human Genetics (University of Würzburg, Germany), especially S. Rost for granting us access to the Illumina MiSeq device. Further thanks go to F. Förster (Department of Bioinformatics, University of Würzburg, Germany) for help with the ITS2 database update. We also thank P. Schloss (University of Michigan, USA) for initial advice on primer design.

Compliance with ethical guidelines

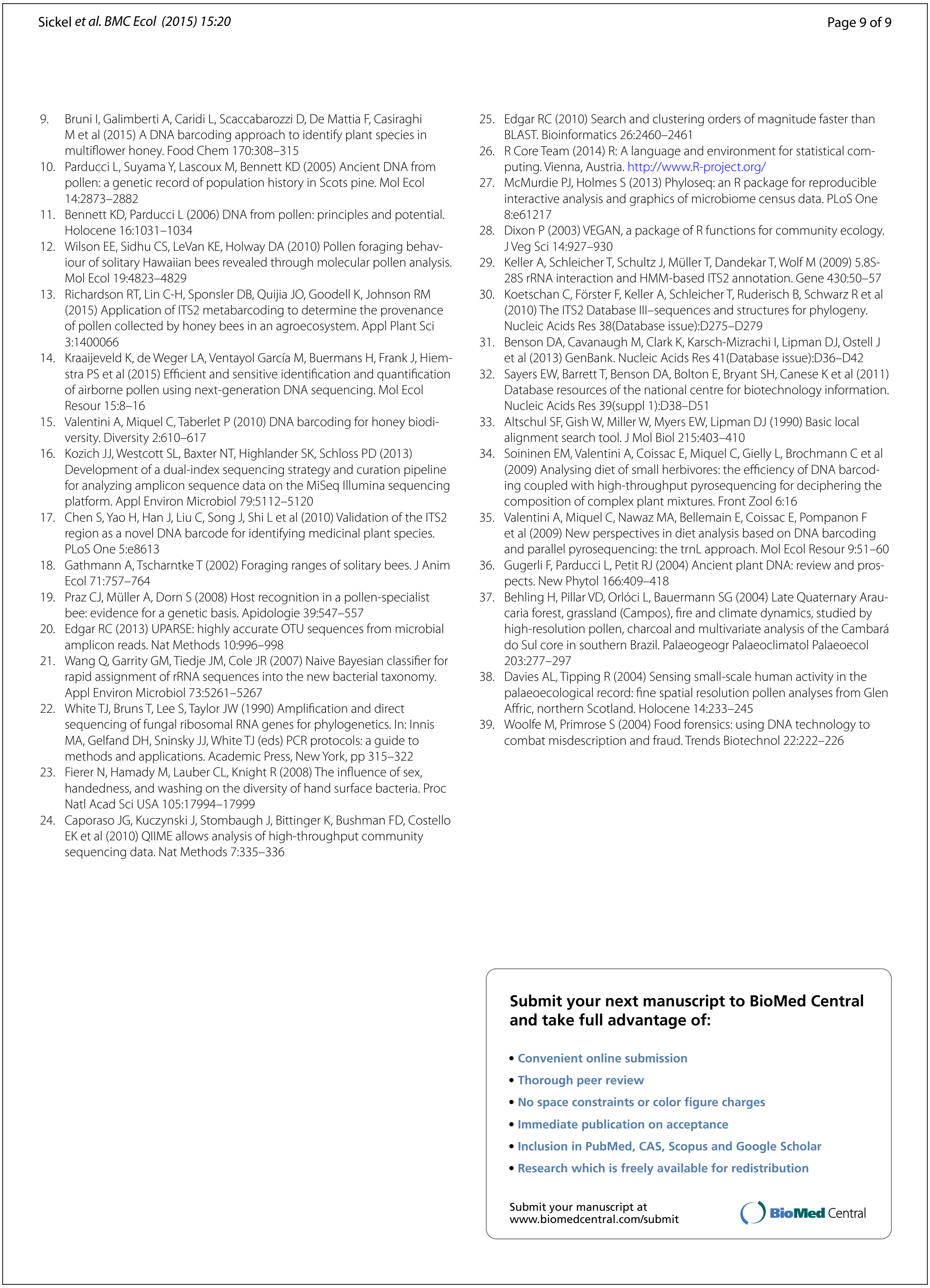
Competing interests

The authors declare that they have no competing interests.

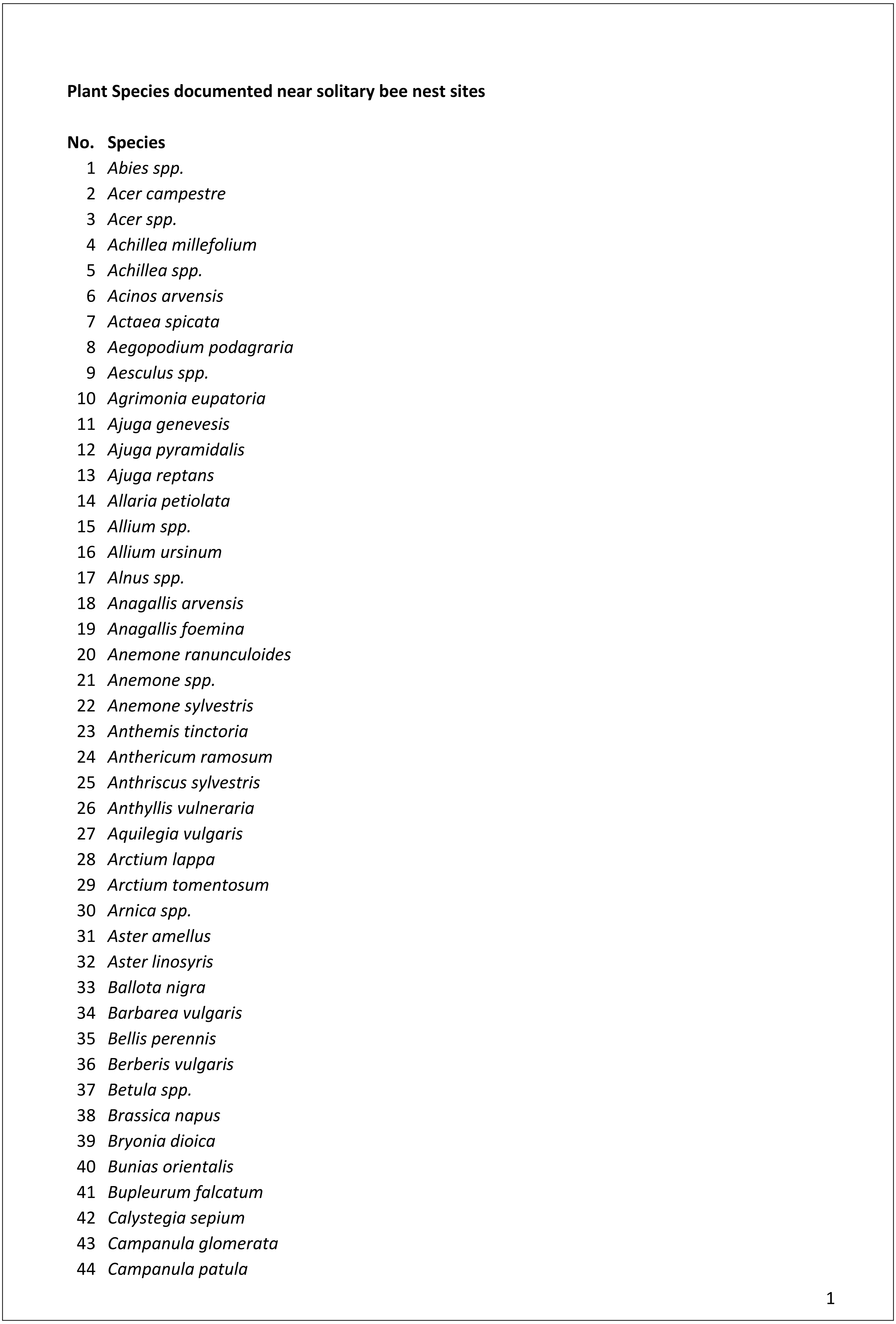
Received: 15 April 2015 Accepted: 25 June 2015
Published online: 22 July 2015

References

1. Carvell C, Westrich P, Meek WR, Pywell RF, Nowakowski M (2006) Assessing the value of annual and perennial forage mixtures for bumblebees by direct observation and pollen analysis. Apidologie 37:326–340
2. Köppler K, Vorwohl G, Koeniger N (2007) Comparison of pollen spectra collected by four different subspecies of the honey bee Apis mellifera. Apidologie 38:341–353
3. Behl M, Horn H, Schwabe A (2008) Analysis of pollen loads in a wild bee community (Hymenoptera: Apidae)—a method for elucidating habitat use and foraging distances. Apidologie 39:456–467
4. Williams NM, Kremen C (2007) Resource distributions among habitats determine solitary bee offspring production in a mosaic landscape. Ecol Appl 17:910–921
5. Krupke CH, Hunt GJ, Eitzer BD, Andino G, Given K (2012) Multiple routes of pesticide exposure for honey bees living near agricultural fields. PLoS One 7:e29268
6. Mullins J, Emberlin J (1997) Sampling pollens. J Aerosol Sci 28:365–370
7. Keller A, Danner N, Grimmer G, Ankenbrand M, von der Ohe K, von der Ohe W et al (2015) Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. Plant Biol 17:558–566
8. Galimberti A, De Mattia F, Bruni I, Scaccabarozzi D, Sandionigi A, Barbuto M et al (2014) A DNA barcoding approach to characterize pollen collected by honeybees. PLoS One 9:e109363

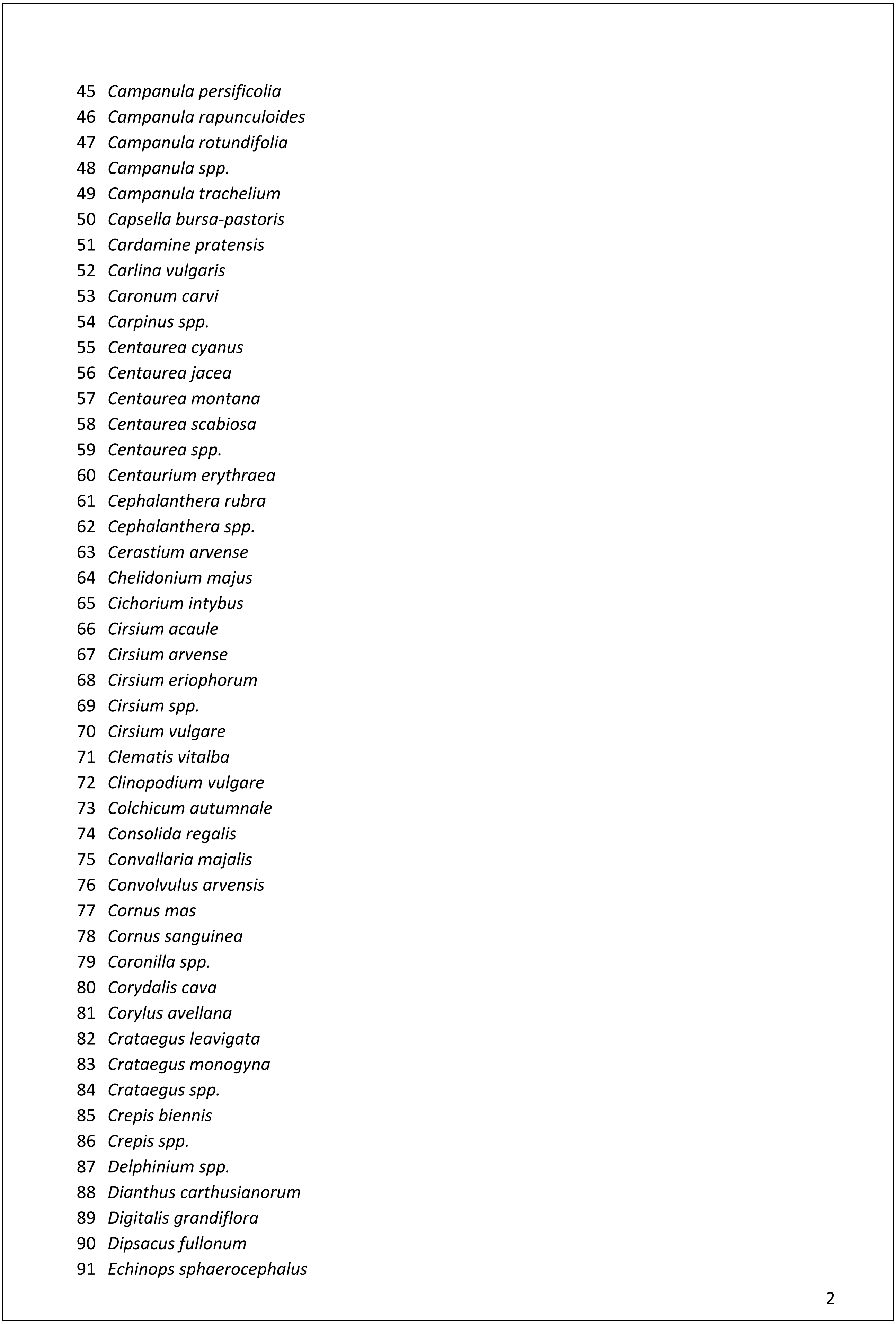


9. Bruni I, Galimberti A, Caridi L, Scaccabarozzi D, De Mattia F, Casiraghi M et al (2015) A DNA barcoding approach to identify plant species in multiflower honey. Food Chem 170:308–315
10. Parducci L, Suyama Y, Lascoux M, Bennett KD (2005) Ancient DNA from pollen: a genetic record of population history in Scots pine. Mol Ecol 14:2873–2882
11. Bennett KD, Parducci L (2006) DNA from pollen: principles and potential. Holocene 16:1031–1034
12. Wilson EE, Sidhu CS, LeVan KE, Holway DA (2010) Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. Mol Ecol 19:4823–4829
13. Richardson RT, Lin C-H, Sponsler DB, Quijia JO, Goodell K, Johnson RM (2015) Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. Appl Plant Sci 3:1400066
14. Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS et al (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Mol Ecol Resour 15:8–16
15. Valentini A, Miquel C, Taberlet P (2010) DNA barcoding for honey biodiversity. Diversity 2:610–617
16. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–5120
17. Chen S, Yao H, Han J, Liu C, Song J, Shi L et al (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One 5:e8613
18. Gathmann A, Tscharntke T (2002) Foraging ranges of solitary bees. J Anim Ecol 71:757–764
19. Praz CJ, Müller A, Dorn S (2008) Host recognition in a pollen-specialist bee: evidence for a genetic basis. Apidologie 39:547–557
20. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998
21. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267
22. White TJ, Bruns T, Lee S, Taylor JW (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols: a guide to methods and applications. Academic Press, New York, pp 315–322
23. Fierer N, Hamady M, Lauber CL, Knight R (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci USA 105:17994–17999
24. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336
25. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461
26. R Core Team (2014) R: A language and environment for statistical computing. Vienna, Austria. http://www.R-project.org/
27. McMurdie PJ, Holmes S (2013) Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217
28. Dixon P (2003) VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930
29. Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M (2009) 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. Gene 430:50–57
30. Koetschan C, Förster F, Keller A, Schleicher T, Ruderisch B, Schwarz R et al (2010) The ITS2 Database III–sequences and structures for phylogeny. Nucleic Acids Res 38(Database issue):D275–D279
31. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J et al (2013) GenBank. Nucleic Acids Res 41(Database issue):D36–D42
32. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K et al (2011) Database resources of the national centre for biotechnology information. Nucleic Acids Res 39(suppl 1):D38–D51
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
34. Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C et al (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Front Zool 6:16
35. Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F et al (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Mol Ecol Resour 9:51–60
36. Gugerli F, Parducci L, Petit RJ (2004) Ancient plant DNA: review and prospects. New Phytol 166:409–418
37. Behling H, Pillar VD, Orlóci L, Bauermann SG (2004) Late Quaternary Araucaria forest, grassland (Campos), fire and climate dynamics, studied by high-resolution pollen, charcoal and multivariate analysis of the Cambará do Sul core in southern Brazil. Palaeogeogr Palaeoclimatol Palaeoecol 203:277–297
38. Davies AL, Tipping R (2004) Sensing small-scale human activity in the palaeoecological record: fine spatial resolution pollen analyses from Glen Affric, northern Scotland. Holocene 14:233–245
39. Woolfe M, Primrose S (2004) Food forensics: using DNA technology to combat misdescription and fraud. Trends Biotechnol 22:222–226

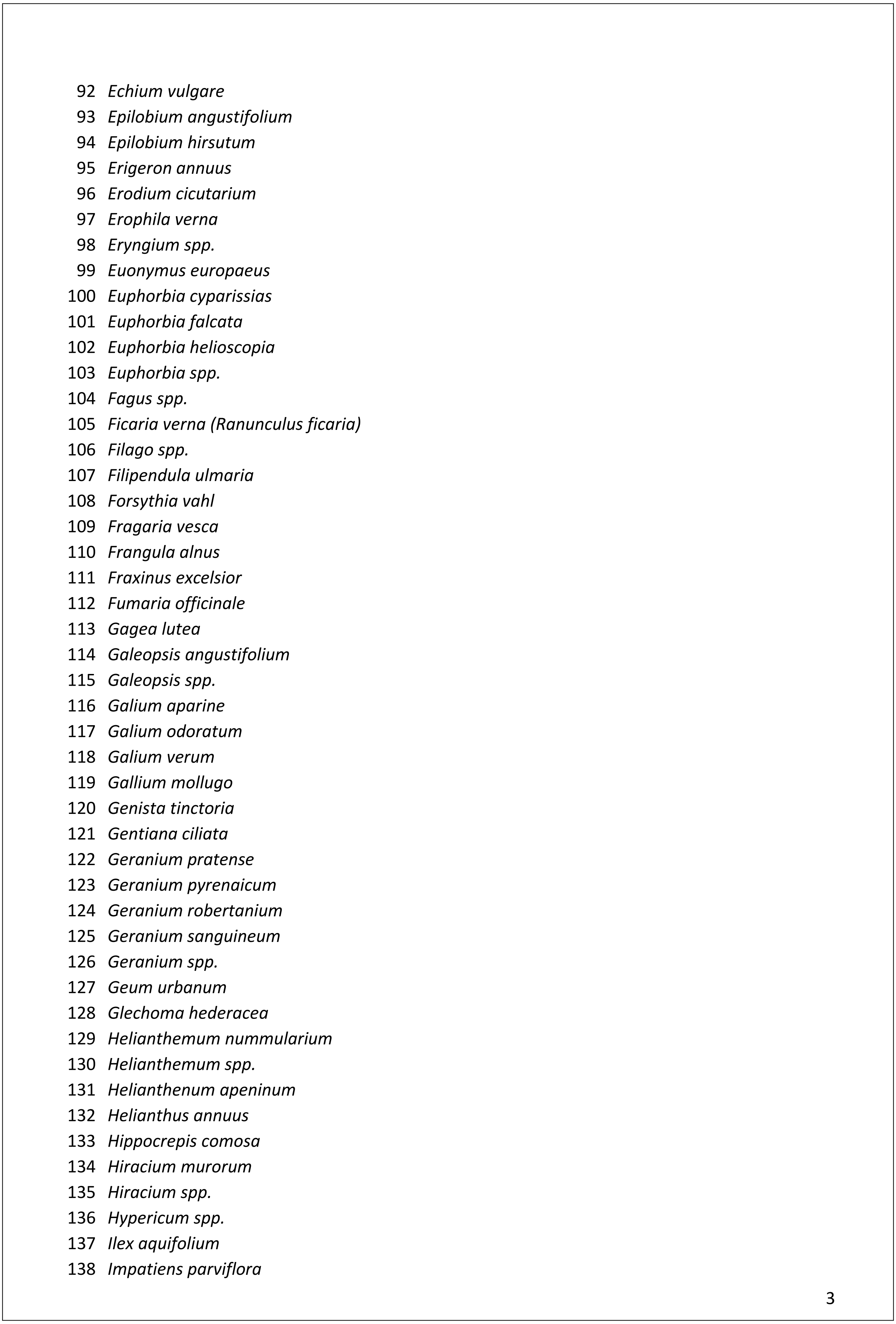


Plant Species documented near solitary bee nest sites

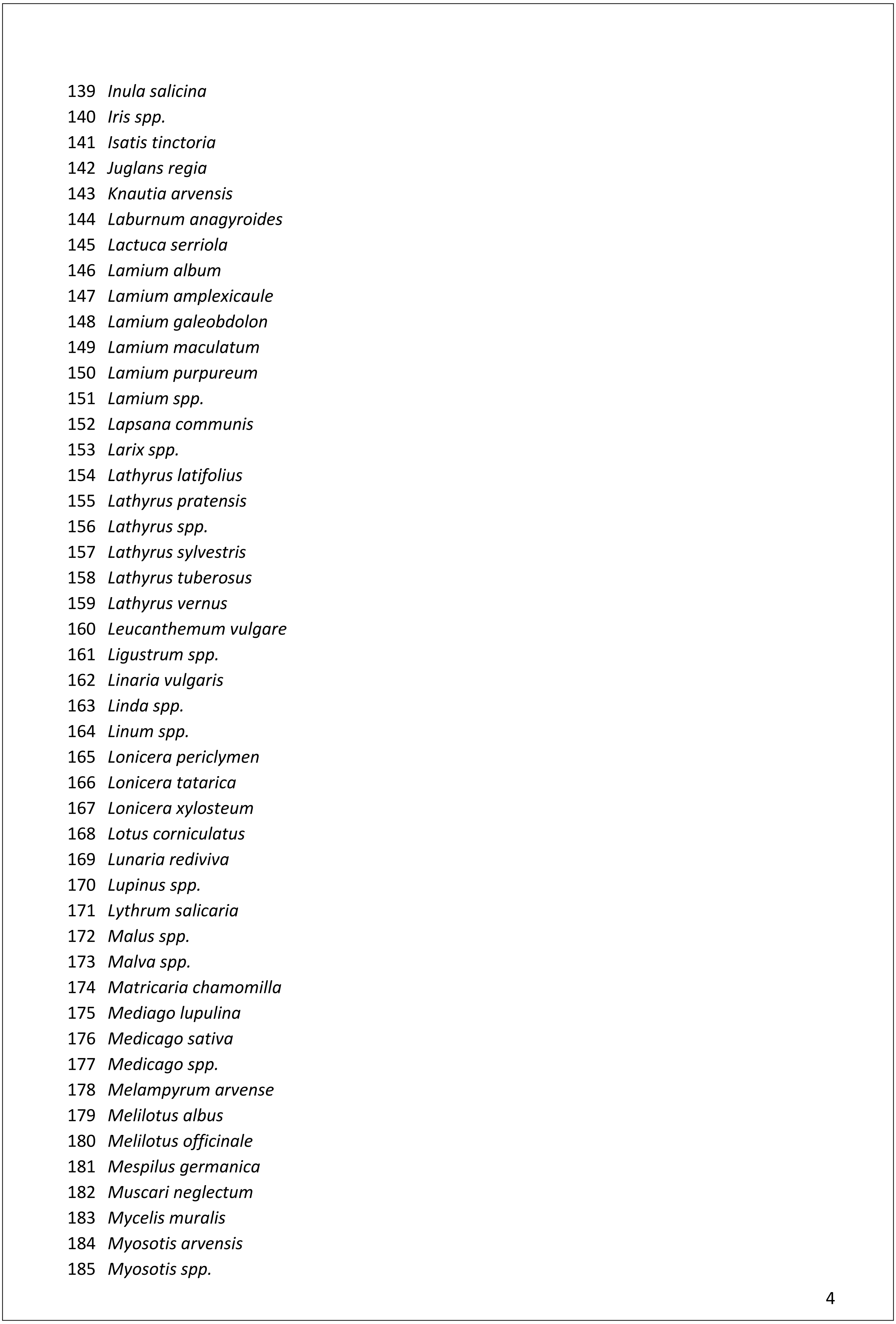
No.	Species
1	*Abies spp.*
2	*Acer campestre*
3	*Acer spp.*
4	*Achillea millefolium*
5	*Achillea spp.*
6	*Acinos arvensis*
7	*Actaea spicata*
8	*Aegopodium podagraria*
9	*Aesculus spp.*
10	*Agrimonia eupatoria*
11	*Ajuga genevesis*
12	*Ajuga pyramidalis*
13	*Ajuga reptans*
14	*Allaria petiolata*
15	*Allium spp.*
16	*Allium ursinum*
17	*Alnus spp.*
18	*Anagallis arvensis*
19	*Anagallis foemina*
20	*Anemone ranunculoides*
21	*Anemone spp.*
22	*Anemone sylvestris*
23	*Anthemis tinctoria*
24	*Anthericum ramosum*
25	*Anthriscus sylvestris*
26	*Anthyllis vulneraria*
27	*Aquilegia vulgaris*
28	*Arctium lappa*
29	*Arctium tomentosum*
30	*Arnica spp.*
31	*Aster amellus*
32	*Aster linosyris*
33	*Ballota nigra*
34	*Barbarea vulgaris*
35	*Bellis perennis*
36	*Berberis vulgaris*
37	*Betula spp.*
38	*Brassica napus*
39	*Bryonia dioica*
40	*Bunias orientalis*
41	*Bupleurum falcatum*
42	*Calystegia sepium*
43	*Campanula glomerata*
44	*Campanula patula*



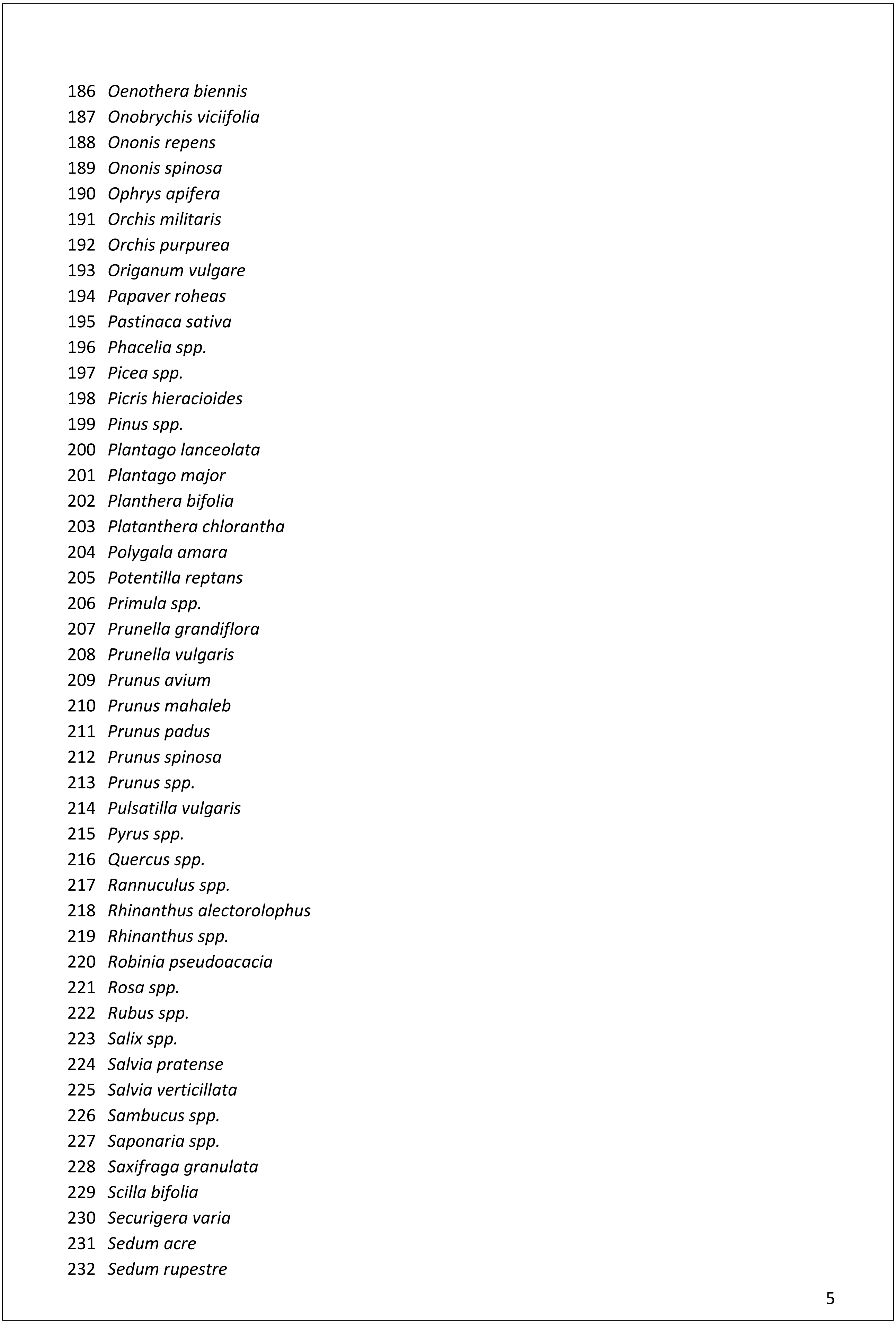
45 *Campanula persificolia*
46 *Campanula rapunculoides*
47 *Campanula rotundifolia*
48 *Campanula spp.*
49 *Campanula trachelium*
50 *Capsella bursa-pastoris*
51 *Cardamine pratensis*
52 *Carlina vulgaris*
53 *Caronum carvi*
54 *Carpinus spp.*
55 *Centaurea cyanus*
56 *Centaurea jacea*
57 *Centaurea montana*
58 *Centaurea scabiosa*
59 *Centaurea spp.*
60 *Centaurium erythraea*
61 *Cephalanthera rubra*
62 *Cephalanthera spp.*
63 *Cerastium arvense*
64 *Chelidonium majus*
65 *Cichorium intybus*
66 *Cirsium acaule*
67 *Cirsium arvense*
68 *Cirsium eriophorum*
69 *Cirsium spp.*
70 *Cirsium vulgare*
71 *Clematis vitalba*
72 *Clinopodium vulgare*
73 *Colchicum autumnale*
74 *Consolida regalis*
75 *Convallaria majalis*
76 *Convolvulus arvensis*
77 *Cornus mas*
78 *Cornus sanguinea*
79 *Coronilla spp.*
80 *Corydalis cava*
81 *Corylus avellana*
82 *Crataegus leavigata*
83 *Crataegus monogyna*
84 *Crataegus spp.*
85 *Crepis biennis*
86 *Crepis spp.*
87 *Delphinium spp.*
88 *Dianthus carthusianorum*
89 *Digitalis grandiflora*
90 *Dipsacus fullonum*
91 *Echinops sphaerocephalus*



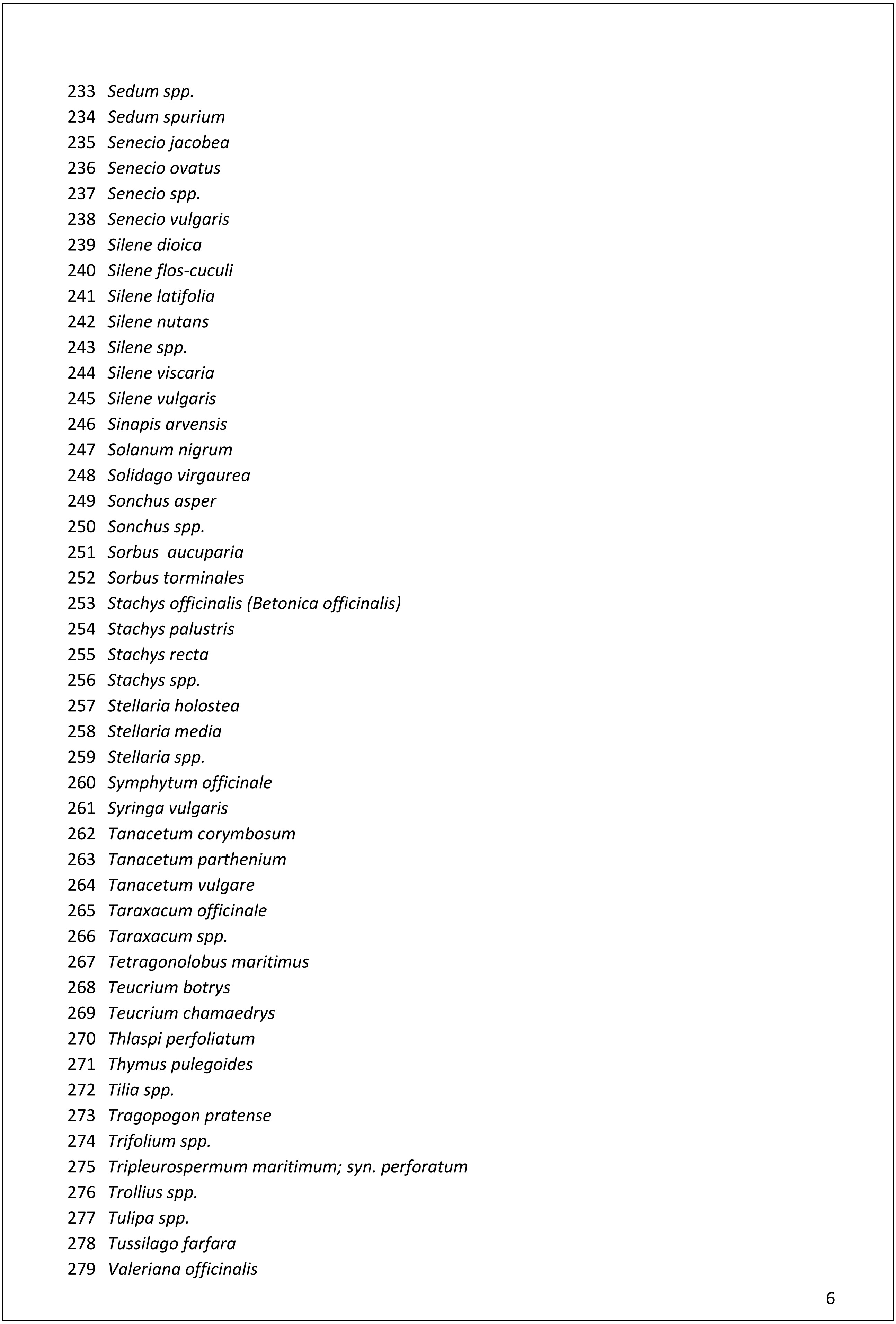
92 *Echium vulgare*
93 *Epilobium angustifolium*
94 *Epilobium hirsutum*
95 *Erigeron annuus*
96 *Erodium cicutarium*
97 *Erophila verna*
98 *Eryngium spp.*
99 *Euonymus europaeus*
100 *Euphorbia cyparissias*
101 *Euphorbia falcata*
102 *Euphorbia helioscopia*
103 *Euphorbia spp.*
104 *Fagus spp.*
105 *Ficaria verna (Ranunculus ficaria)*
106 *Filago spp.*
107 *Filipendula ulmaria*
108 *Forsythia vahl*
109 *Fragaria vesca*
110 *Frangula alnus*
111 *Fraxinus excelsior*
112 *Fumaria officinale*
113 *Gagea lutea*
114 *Galeopsis angustifolium*
115 *Galeopsis spp.*
116 *Galium aparine*
117 *Galium odoratum*
118 *Galium verum*
119 *Gallium mollugo*
120 *Genista tinctoria*
121 *Gentiana ciliata*
122 *Geranium pratense*
123 *Geranium pyrenaicum*
124 *Geranium robertanium*
125 *Geranium sanguineum*
126 *Geranium spp.*
127 *Geum urbanum*
128 *Glechoma hederacea*
129 *Helianthemum nummularium*
130 *Helianthemum spp.*
131 *Helianthenum apeninum*
132 *Helianthus annuus*
133 *Hippocrepis comosa*
134 *Hiracium murorum*
135 *Hiracium spp.*
136 *Hypericum spp.*
137 *Ilex aquifolium*
138 *Impatiens parviflora*



139 *Inula salicina*
140 *Iris spp.*
141 *Isatis tinctoria*
142 *Juglans regia*
143 *Knautia arvensis*
144 *Laburnum anagyroides*
145 *Lactuca serriola*
146 *Lamium album*
147 *Lamium amplexicaule*
148 *Lamium galeobdolon*
149 *Lamium maculatum*
150 *Lamium purpureum*
151 *Lamium spp.*
152 *Lapsana communis*
153 *Larix spp.*
154 *Lathyrus latifolius*
155 *Lathyrus pratensis*
156 *Lathyrus spp.*
157 *Lathyrus sylvestris*
158 *Lathyrus tuberosus*
159 *Lathyrus vernus*
160 *Leucanthemum vulgare*
161 *Ligustrum spp.*
162 *Linaria vulgaris*
163 *Linda spp.*
164 *Linum spp.*
165 *Lonicera periclymen*
166 *Lonicera tatarica*
167 *Lonicera xylosteum*
168 *Lotus corniculatus*
169 *Lunaria rediviva*
170 *Lupinus spp.*
171 *Lythrum salicaria*
172 *Malus spp.*
173 *Malva spp.*
174 *Matricaria chamomilla*
175 *Mediago lupulina*
176 *Medicago sativa*
177 *Medicago spp.*
178 *Melampyrum arvense*
179 *Melilotus albus*
180 *Melilotus officinale*
181 *Mespilus germanica*
182 *Muscari neglectum*
183 *Mycelis muralis*
184 *Myosotis arvensis*
185 *Myosotis spp.*



186 *Oenothera biennis*
187 *Onobrychis viciifolia*
188 *Ononis repens*
189 *Ononis spinosa*
190 *Ophrys apifera*
191 *Orchis militaris*
192 *Orchis purpurea*
193 *Origanum vulgare*
194 *Papaver roheas*
195 *Pastinaca sativa*
196 *Phacelia spp.*
197 *Picea spp.*
198 *Picris hieracioides*
199 *Pinus spp.*
200 *Plantago lanceolata*
201 *Plantago major*
202 *Planthera bifolia*
203 *Platanthera chlorantha*
204 *Polygala amara*
205 *Potentilla reptans*
206 *Primula spp.*
207 *Prunella grandiflora*
208 *Prunella vulgaris*
209 *Prunus avium*
210 *Prunus mahaleb*
211 *Prunus padus*
212 *Prunus spinosa*
213 *Prunus spp.*
214 *Pulsatilla vulgaris*
215 *Pyrus spp.*
216 *Quercus spp.*
217 *Rannuculus spp.*
218 *Rhinanthus alectorolophus*
219 *Rhinanthus spp.*
220 *Robinia pseudoacacia*
221 *Rosa spp.*
222 *Rubus spp.*
223 *Salix spp.*
224 *Salvia pratense*
225 *Salvia verticillata*
226 *Sambucus spp.*
227 *Saponaria spp.*
228 *Saxifraga granulata*
229 *Scilla bifolia*
230 *Securigera varia*
231 *Sedum acre*
232 *Sedum rupestre*



233 *Sedum spp.*

234 *Sedum spurium*

235 *Senecio jacobea*

236 *Senecio ovatus*

237 *Senecio spp.*

238 *Senecio vulgaris*

239 *Silene dioica*

240 *Silene flos-cuculi*

241 *Silene latifolia*

242 *Silene nutans*

243 *Silene spp.*

244 *Silene viscaria*

245 *Silene vulgaris*

246 *Sinapis arvensis*

247 *Solanum nigrum*

248 *Solidago virgaurea*

249 *Sonchus asper*

250 *Sonchus spp.*

251 *Sorbus aucuparia*

252 *Sorbus torminales*

253 *Stachys officinalis (Betonica officinalis)*

254 *Stachys palustris*

255 *Stachys recta*

256 *Stachys spp.*

257 *Stellaria holostea*

258 *Stellaria media*

259 *Stellaria spp.*

260 *Symphytum officinale*

261 *Syringa vulgaris*

262 *Tanacetum corymbosum*

263 *Tanacetum parthenium*

264 *Tanacetum vulgare*

265 *Taraxacum officinale*

266 *Taraxacum spp.*

267 *Tetragonolobus maritimus*

268 *Teucrium botrys*

269 *Teucrium chamaedrys*

270 *Thlaspi perfoliatum*

271 *Thymus pulegoides*

272 *Tilia spp.*

273 *Tragopogon pratense*

274 *Trifolium spp.*

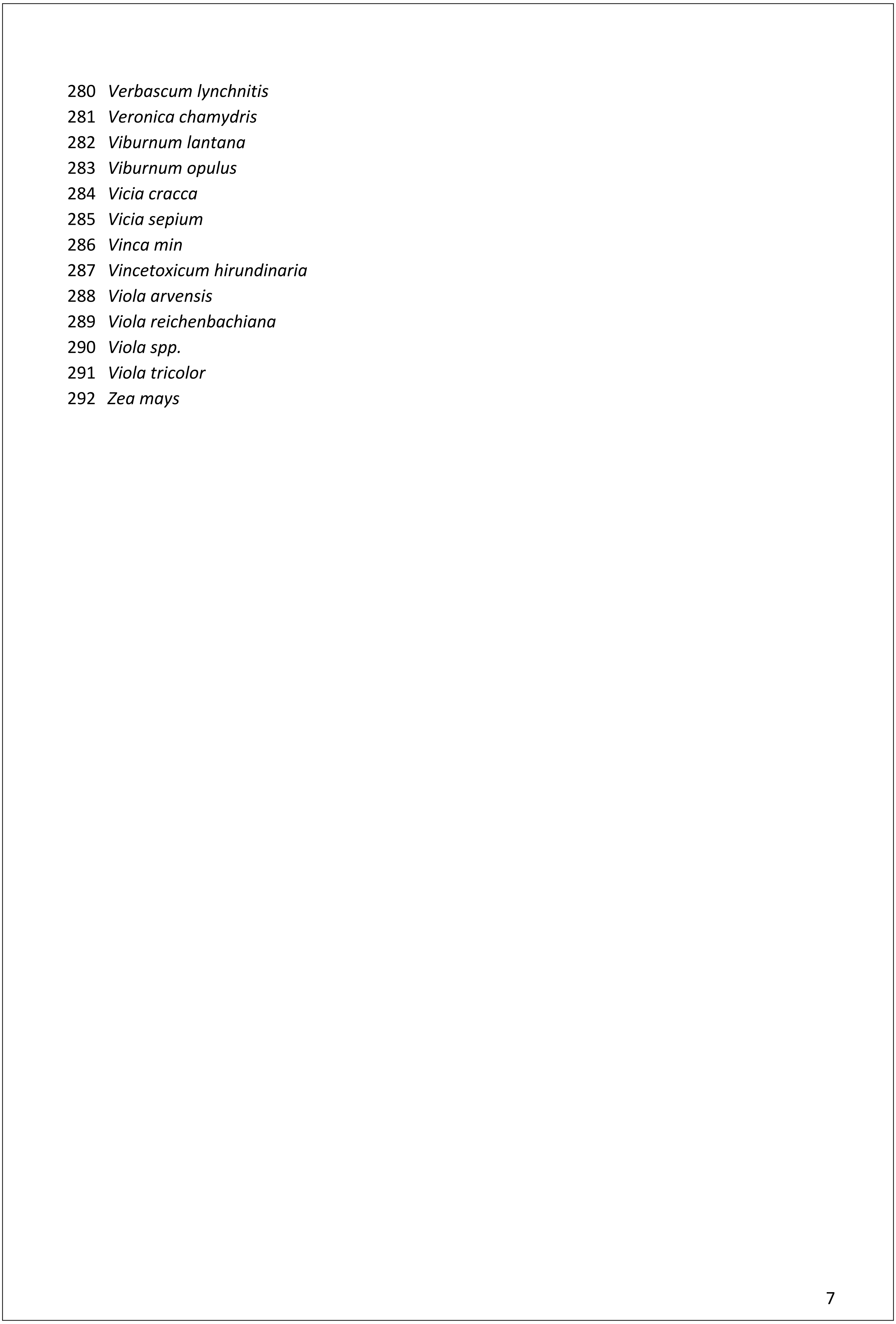
275 *Tripleurospermum maritimum; syn. perforatum*

276 *Trollius spp.*

277 *Tulipa spp.*

278 *Tussilago farfara*

279 *Valeriana officinalis*



280 *Verbascum lynchnitis*
281 *Veronica chamydris*
282 *Viburnum lantana*
283 *Viburnum opulus*
284 *Vicia cracca*
285 *Vicia sepium*
286 *Vinca min*
287 *Vincetoxicum hirundinaria*
288 *Viola arvensis*
289 *Viola reichenbachiana*
290 *Viola spp.*
291 *Viola tricolor*
292 *Zea mays*

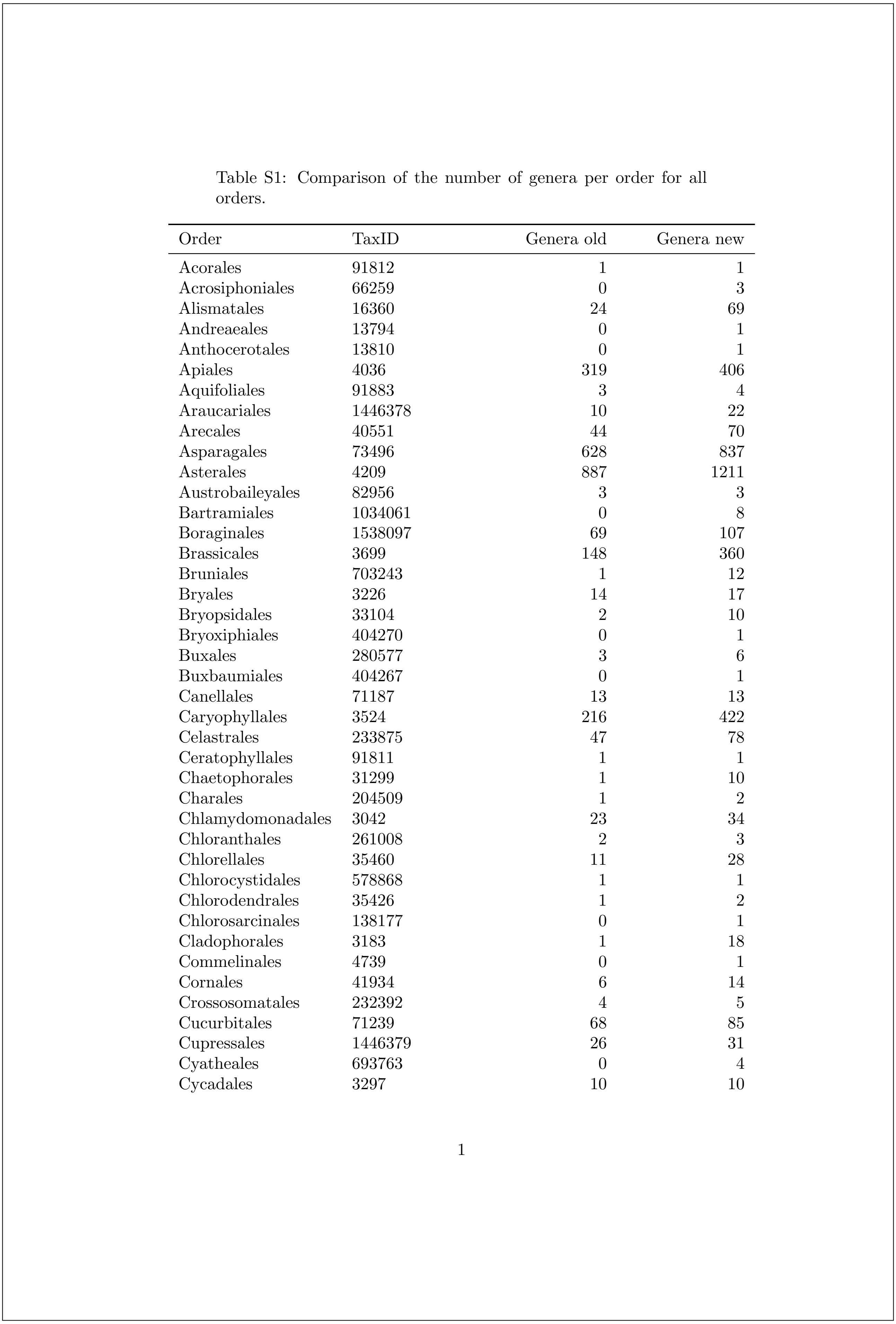


Table S1: Comparison of the number of genera per order for all orders.

Order	TaxID	Genera old	Genera new
Acorales	91812	1	1
Acrosiphoniales	66259	0	3
Alismatales	16360	24	69
Andreaeales	13794	0	1
Anthocerotales	13810	0	1
Apiales	4036	319	406
Aquifoliales	91883	3	4
Araucariales	1446378	10	22
Arecales	40551	44	70
Asparagales	73496	628	837
Asterales	4209	887	1211
Austrobaileyales	82956	3	3
Bartramiales	1034061	0	8
Boraginales	1538097	69	107
Brassicales	3699	148	360
Bruniales	703243	1	12
Bryales	3226	14	17
Bryopsidales	33104	2	10
Bryoxiphiales	404270	0	1
Buxales	280577	3	6
Buxbaumiales	404267	0	1
Canellales	71187	13	13
Caryophyllales	3524	216	422
Celastrales	233875	47	78
Ceratophyllales	91811	1	1
Chaetophorales	31299	1	10
Charales	204509	1	2
Chlamydomonadales	3042	23	34
Chloranthales	261008	2	3
Chlorellales	35460	11	28
Chlorocystidales	578868	1	1
Chlorodendrales	35426	1	2
Chlorosarcinales	138177	0	1
Cladophorales	3183	1	18
Commelinales	4739	0	1
Cornales	41934	6	14
Crossosomatales	232392	4	5
Cucurbitales	71239	68	85
Cupressales	1446379	26	31
Cyatheales	693763	0	4
Cycadales	3297	10	10

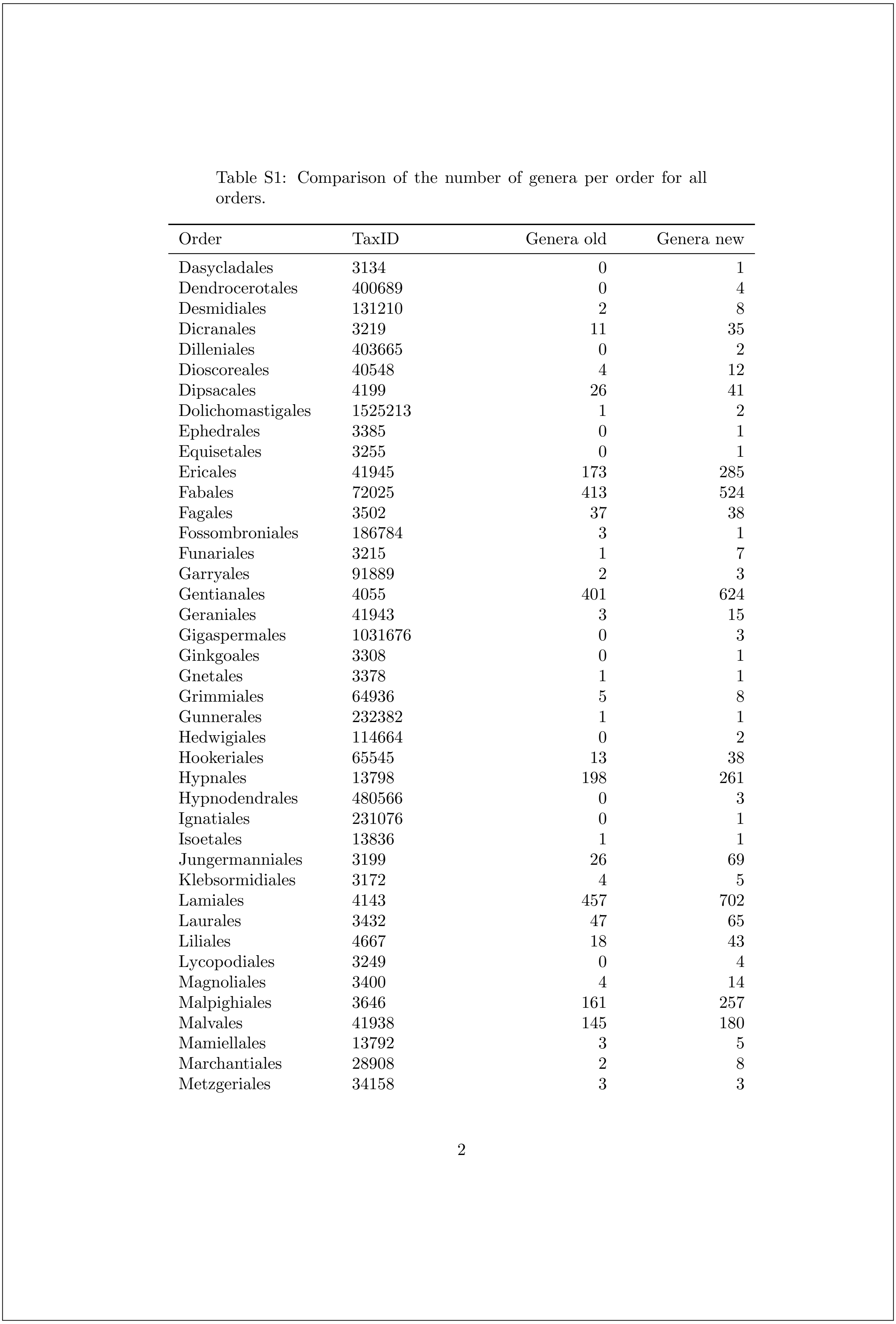


Table S1: Comparison of the number of genera per order for all orders.

Order	TaxID	Genera old	Genera new
Dasycladales	3134	0	1
Dendrocerotales	400689	0	4
Desmidiales	131210	2	8
Dicranales	3219	11	35
Dilleniales	403665	0	2
Dioscoreales	40548	4	12
Dipsacales	4199	26	41
Dolichomastigales	1525213	1	2
Ephedrales	3385	0	1
Equisetales	3255	0	1
Ericales	41945	173	285
Fabales	72025	413	524
Fagales	3502	37	38
Fossombroniales	186784	3	1
Funariales	3215	1	7
Garryales	91889	2	3
Gentianales	4055	401	624
Geraniales	41943	3	15
Gigaspermales	1031676	0	3
Ginkgoales	3308	0	1
Gnetales	3378	1	1
Grimmiales	64936	5	8
Gunnerales	232382	1	1
Hedwigiales	114664	0	2
Hookeriales	65545	13	38
Hypnales	13798	198	261
Hypnodendrales	480566	0	3
Ignatiales	231076	0	1
Isoetales	13836	1	1
Jungermanniales	3199	26	69
Klebsormidiales	3172	4	5
Lamiales	4143	457	702
Laurales	3432	47	65
Liliales	4667	18	43
Lycopodiales	3249	0	4
Magnoliales	3400	4	14
Malpighiales	3646	161	257
Malvales	41938	145	180
Mamiellales	13792	3	5
Marchantiales	28908	2	8
Metzgeriales	34158	3	3

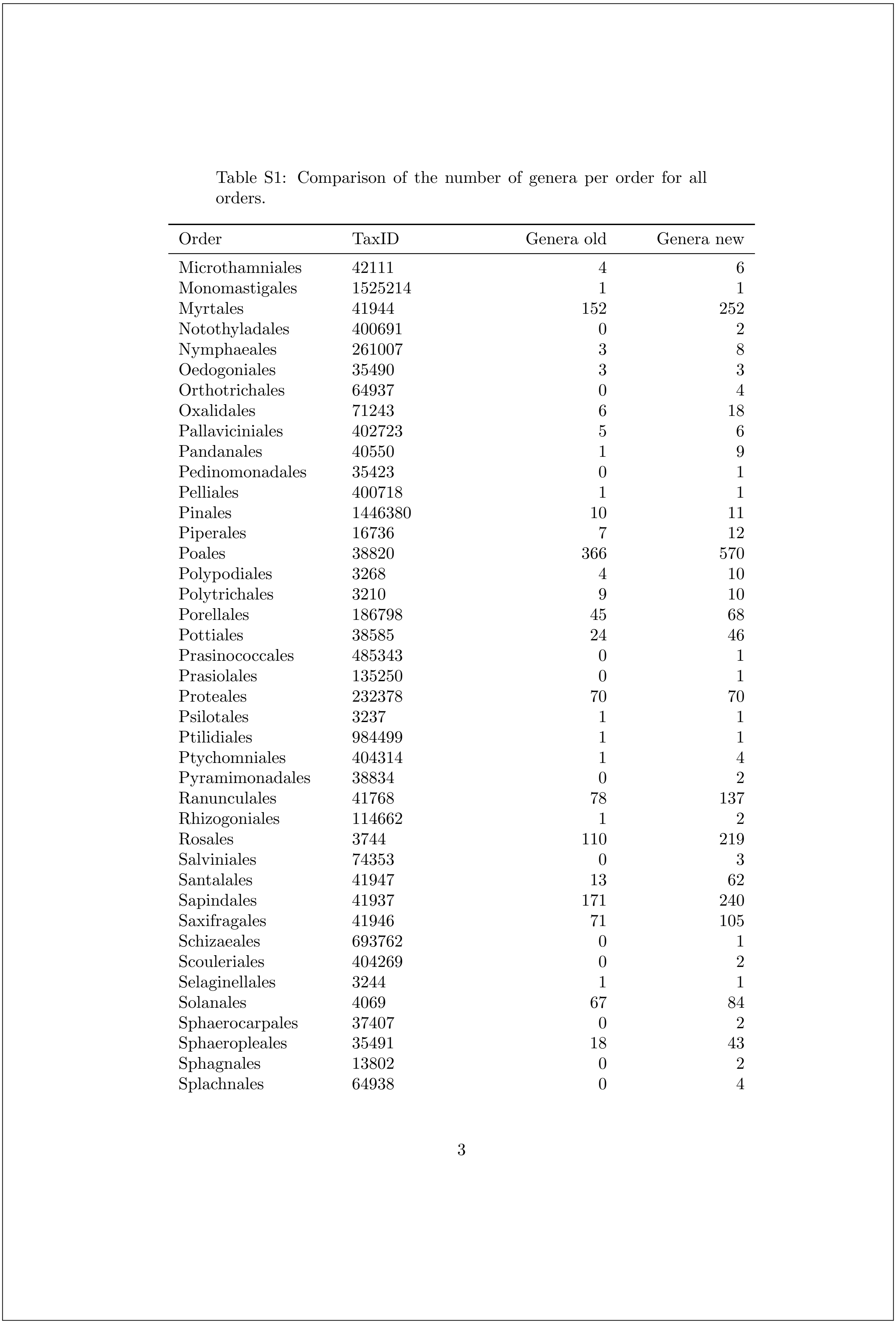


Table S1: Comparison of the number of genera per order for all orders.

Order	TaxID	Genera old	Genera new
Microthamniales	42111	4	6
Monomastigales	1525214	1	1
Myrtales	41944	152	252
Notothyladales	400691	0	2
Nymphaeales	261007	3	8
Oedogoniales	35490	3	3
Orthotrichales	64937	0	4
Oxalidales	71243	6	18
Pallaviciniales	402723	5	6
Pandanales	40550	1	9
Pedinomonadales	35423	0	1
Pelliales	400718	1	1
Pinales	1446380	10	11
Piperales	16736	7	12
Poales	38820	366	570
Polypodiales	3268	4	10
Polytrichales	3210	9	10
Porellales	186798	45	68
Pottiales	38585	24	46
Prasinococcales	485343	0	1
Prasiolales	135250	0	1
Proteales	232378	70	70
Psilotales	3237	1	1
Ptilidiales	984499	1	1
Ptychomniales	404314	1	4
Pyramimonadales	38834	0	2
Ranunculales	41768	78	137
Rhizogoniales	114662	1	2
Rosales	3744	110	219
Salviniales	74353	0	3
Santalales	41947	13	62
Sapindales	41937	171	240
Saxifragales	41946	71	105
Schizaeales	693762	0	1
Scouleriales	404269	0	2
Selaginellales	3244	1	1
Solanales	4069	67	84
Sphaerocarpales	37407	0	2
Sphaeropleales	35491	18	43
Sphagnales	13802	0	2
Splachnales	64938	0	4

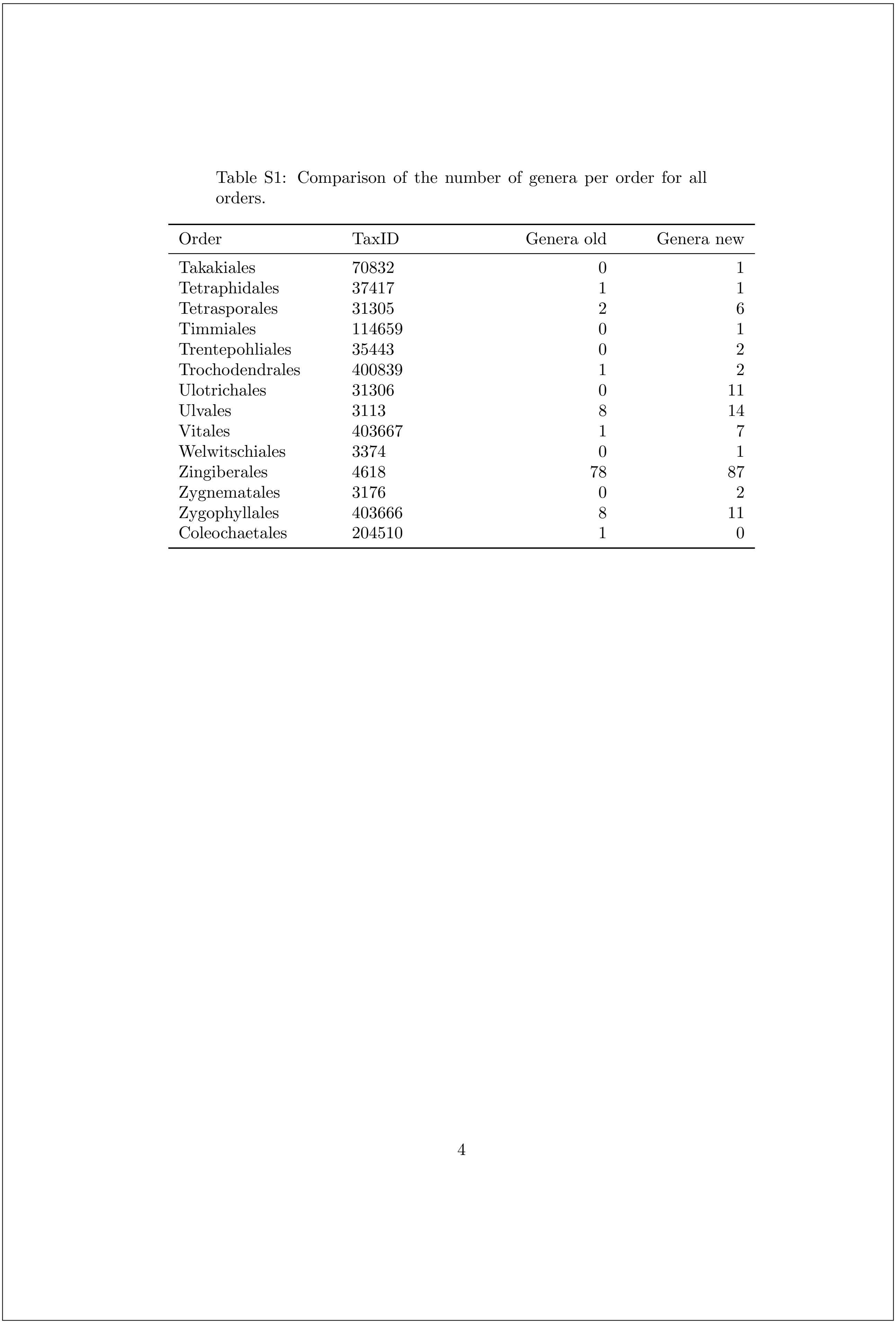


Table S1: Comparison of the number of genera per order for all orders.

Order	TaxID	Genera old	Genera new
Takakiales	70832	0	1
Tetraphidales	37417	1	1
Tetrasporales	31305	2	6
Timmiales	114659	0	1
Trentepohliales	35443	0	2
Trochodendrales	400839	1	2
Ulotrichales	31306	0	11
Ulvales	3113	8	14
Vitales	403667	1	7
Welwitschiales	3374	0	1
Zingiberales	4618	78	87
Zygnematales	3176	0	2
Zygophyllales	403666	8	11
Coleochaetales	204510	1	0

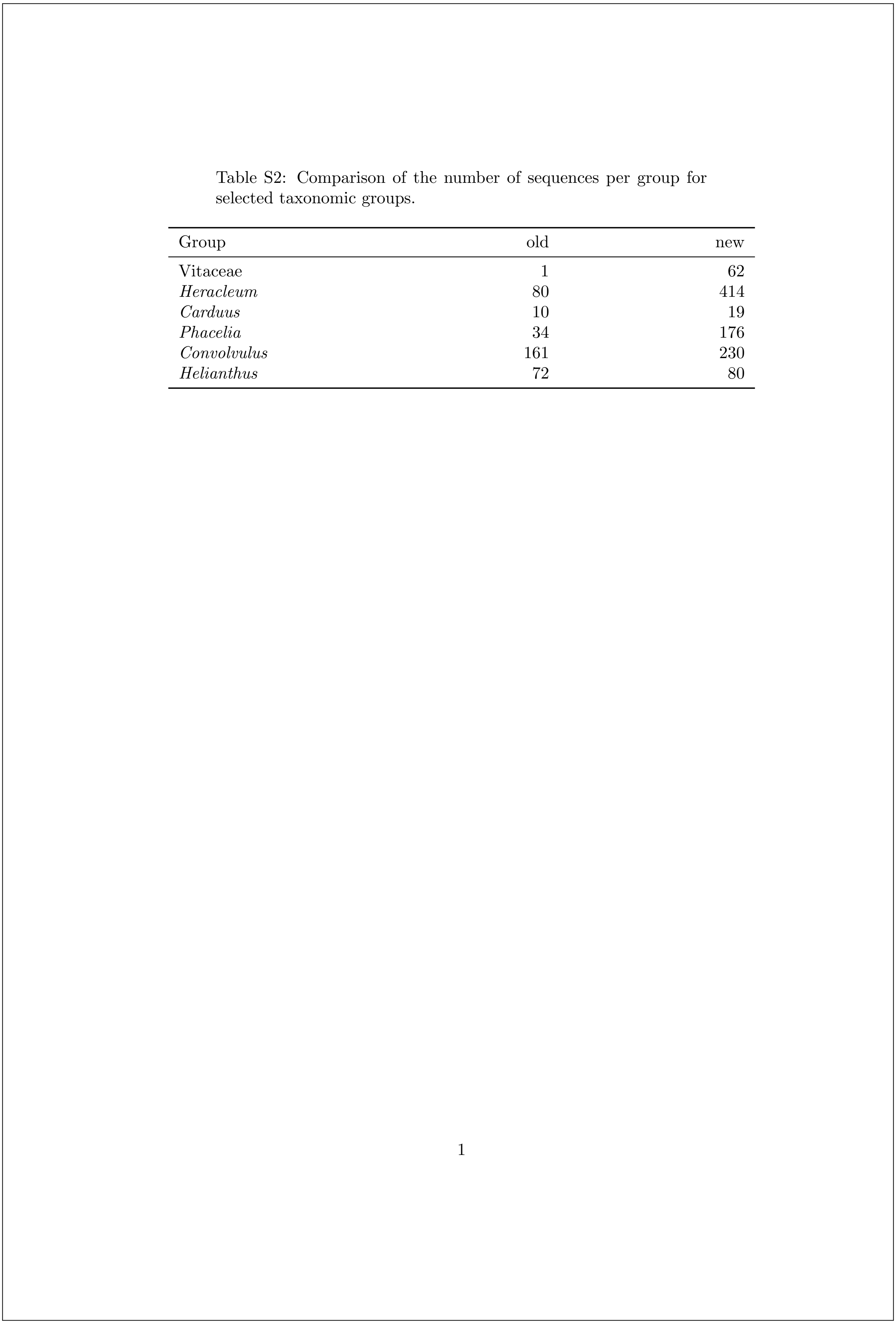


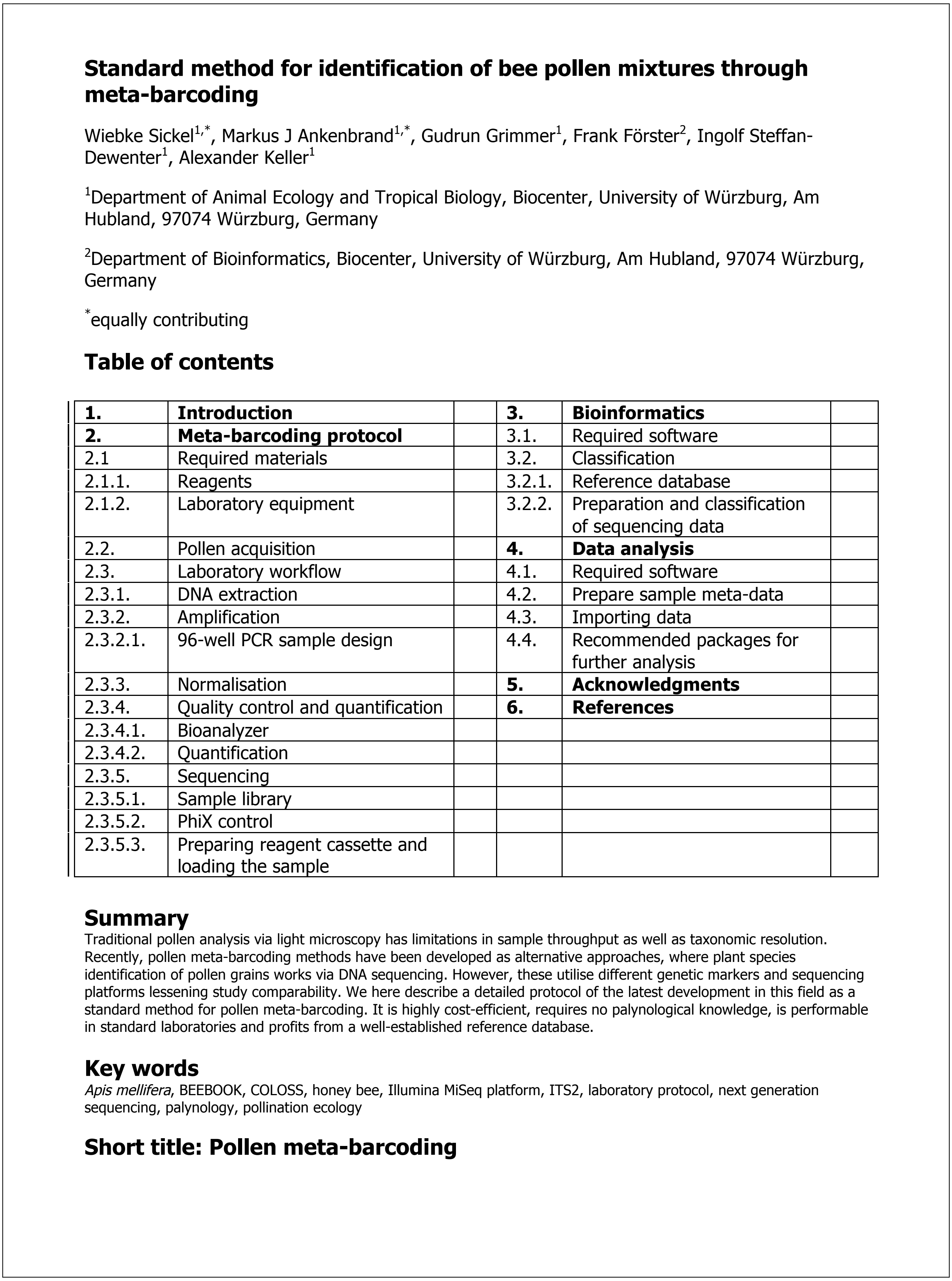
Table S2: Comparison of the number of sequences per group for selected taxonomic groups.

Group	old	new
Vitaceae	1	62
Heracleum	80	414
Carduus	10	19
Phacelia	34	176
Convolvulus	161	230
Helianthus	72	80

5.4 STANDARD METHOD FOR IDENTIFICATION OF BEE POLLEN MIXTURES THROUGH META-BARCODING

– accepted in *COLOSS BEEBOOK*
(in press) –

This publication is also part of Sickel (2017). It is part of both theses because of complementary contributions.



Standard method for identification of bee pollen mixtures through meta-barcoding

Wiebke Sickel[1,*], Markus J Ankenbrand[1,*], Gudrun Grimmer[1], Frank Förster[2], Ingolf Steffan-Dewenter[1], Alexander Keller[1]

[1]Department of Animal Ecology and Tropical Biology, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

[2]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

[*]equally contributing

Table of contents

1.	**Introduction**		**3.**	**Bioinformatics**	
2.	**Meta-barcoding protocol**		3.1.	Required software	
2.1	Required materials		3.2.	Classification	
2.1.1.	Reagents		3.2.1.	Reference database	
2.1.2.	Laboratory equipment		3.2.2.	Preparation and classification of sequencing data	
2.2.	Pollen acquisition		**4.**	**Data analysis**	
2.3.	Laboratory workflow		4.1.	Required software	
2.3.1.	DNA extraction		4.2.	Prepare sample meta-data	
2.3.2.	Amplification		4.3.	Importing data	
2.3.2.1.	96-well PCR sample design		4.4.	Recommended packages for further analysis	
2.3.3.	Normalisation		**5.**	**Acknowledgments**	
2.3.4.	Quality control and quantification		**6.**	**References**	
2.3.4.1.	Bioanalyzer				
2.3.4.2.	Quantification				
2.3.5.	Sequencing				
2.3.5.1.	Sample library				
2.3.5.2.	PhiX control				
2.3.5.3.	Preparing reagent cassette and loading the sample				

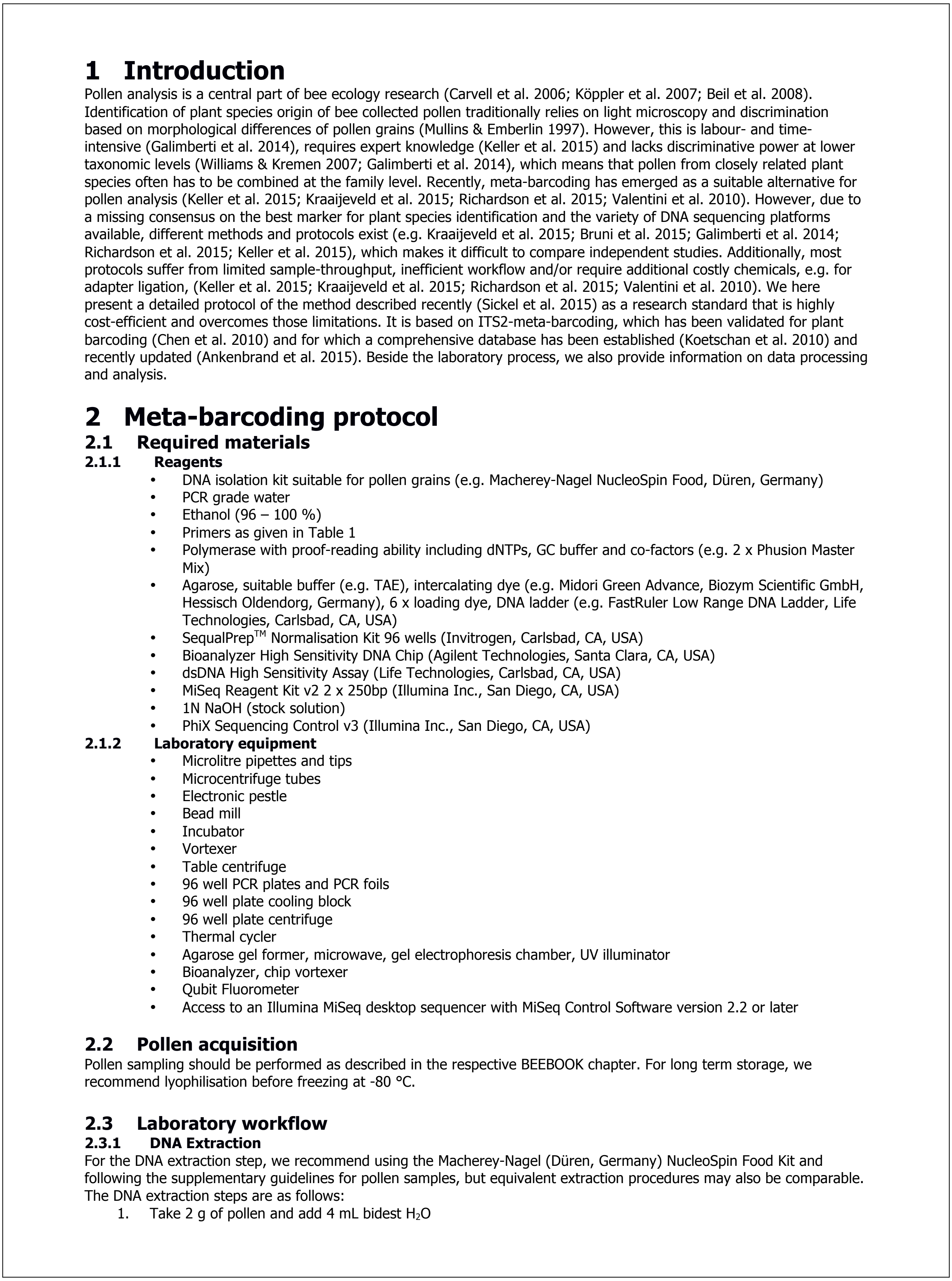
Summary

Traditional pollen analysis via light microscopy has limitations in sample throughput as well as taxonomic resolution. Recently, pollen meta-barcoding methods have been developed as alternative approaches, where plant species identification of pollen grains works via DNA sequencing. However, these utilise different genetic markers and sequencing platforms lessening study comparability. We here describe a detailed protocol of the latest development in this field as a standard method for pollen meta-barcoding. It is highly cost-efficient, requires no palynological knowledge, is performable in standard laboratories and profits from a well-established reference database.

Key words

Apis mellifera, BEEBOOK, COLOSS, honey bee, Illumina MiSeq platform, ITS2, laboratory protocol, next generation sequencing, palynology, pollination ecology

Short title: Pollen meta-barcoding



1 Introduction

Pollen analysis is a central part of bee ecology research (Carvell et al. 2006; Köppler et al. 2007; Beil et al. 2008). Identification of plant species origin of bee collected pollen traditionally relies on light microscopy and discrimination based on morphological differences of pollen grains (Mullins & Emberlin 1997). However, this is labour- and time-intensive (Galimberti et al. 2014), requires expert knowledge (Keller et al. 2015) and lacks discriminative power at lower taxonomic levels (Williams & Kremen 2007; Galimberti et al. 2014), which means that pollen from closely related plant species often has to be combined at the family level. Recently, meta-barcoding has emerged as a suitable alternative for pollen analysis (Keller et al. 2015; Kraaijeveld et al. 2015; Richardson et al. 2015; Valentini et al. 2010). However, due to a missing consensus on the best marker for plant species identification and the variety of DNA sequencing platforms available, different methods and protocols exist (e.g. Kraaijeveld et al. 2015; Bruni et al. 2015; Galimberti et al. 2014; Richardson et al. 2015; Keller et al. 2015), which makes it difficult to compare independent studies. Additionally, most protocols suffer from limited sample-throughput, inefficient workflow and/or require additional costly chemicals, e.g. for adapter ligation, (Keller et al. 2015; Kraaijeveld et al. 2015; Richardson et al. 2015; Valentini et al. 2010). We here present a detailed protocol of the method described recently (Sickel et al. 2015) as a research standard that is highly cost-efficient and overcomes those limitations. It is based on ITS2-meta-barcoding, which has been validated for plant barcoding (Chen et al. 2010) and for which a comprehensive database has been established (Koetschan et al. 2010) and recently updated (Ankenbrand et al. 2015). Beside the laboratory process, we also provide information on data processing and analysis.

2 Meta-barcoding protocol

2.1 Required materials

2.1.1 Reagents

- DNA isolation kit suitable for pollen grains (e.g. Macherey-Nagel NucleoSpin Food, Düren, Germany)
- PCR grade water
- Ethanol (96 – 100 %)
- Primers as given in Table 1
- Polymerase with proof-reading ability including dNTPs, GC buffer and co-factors (e.g. 2 x Phusion Master Mix)
- Agarose, suitable buffer (e.g. TAE), intercalating dye (e.g. Midori Green Advance, Biozym Scientific GmbH, Hessisch Oldendorg, Germany), 6 x loading dye, DNA ladder (e.g. FastRuler Low Range DNA Ladder, Life Technologies, Carlsbad, CA, USA)
- SequalPrep^{TM} Normalisation Kit 96 wells (Invitrogen, Carlsbad, CA, USA)
- Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA, USA)
- dsDNA High Sensitivity Assay (Life Technologies, Carlsbad, CA, USA)
- MiSeq Reagent Kit v2 2 x 250bp (Illumina Inc., San Diego, CA, USA)
- 1N NaOH (stock solution)
- PhiX Sequencing Control v3 (Illumina Inc., San Diego, CA, USA)

2.1.2 Laboratory equipment

- Microlitre pipettes and tips
- Microcentrifuge tubes
- Electronic pestle
- Bead mill
- Incubator
- Vortexer
- Table centrifuge
- 96 well PCR plates and PCR foils
- 96 well plate cooling block
- 96 well plate centrifuge
- Thermal cycler
- Agarose gel former, microwave, gel electrophoresis chamber, UV illuminator
- Bioanalyzer, chip vortexer
- Qubit Fluorometer
- Access to an Illumina MiSeq desktop sequencer with MiSeq Control Software version 2.2 or later

2.2 Pollen acquisition

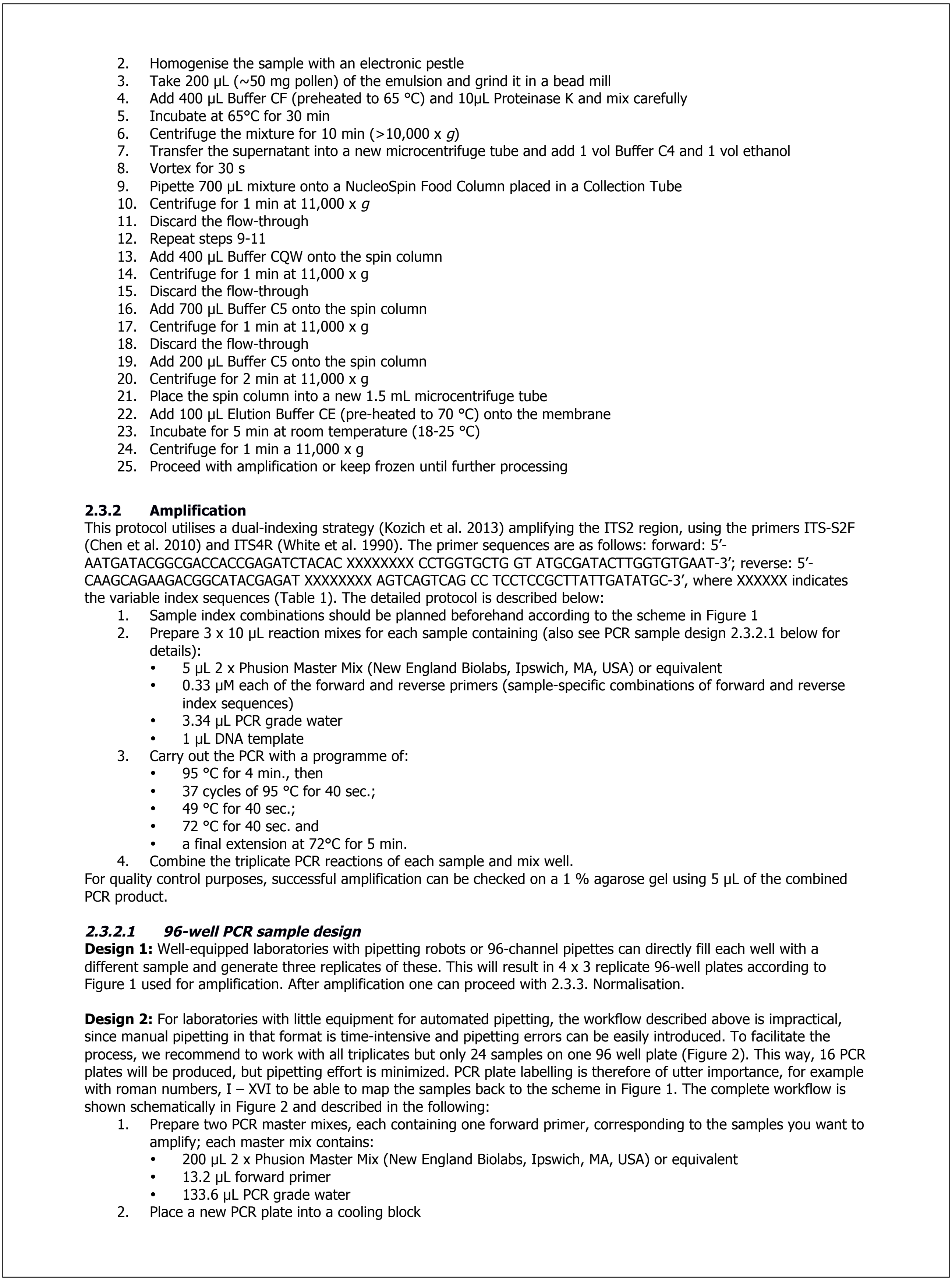
Pollen sampling should be performed as described in the respective BEEBOOK chapter. For long term storage, we recommend lyophilisation before freezing at -80 °C.

2.3 Laboratory workflow

2.3.1 DNA Extraction

For the DNA extraction step, we recommend using the Macherey-Nagel (Düren, Germany) NucleoSpin Food Kit and following the supplementary guidelines for pollen samples, but equivalent extraction procedures may also be comparable. The DNA extraction steps are as follows:

1. Take 2 g of pollen and add 4 mL bidest H_2O



2. Homogenise the sample with an electronic pestle
3. Take 200 µL (~50 mg pollen) of the emulsion and grind it in a bead mill
4. Add 400 µL Buffer CF (preheated to 65 °C) and 10µL Proteinase K and mix carefully
5. Incubate at 65°C for 30 min
6. Centrifuge the mixture for 10 min (>10,000 x *g*)
7. Transfer the supernatant into a new microcentrifuge tube and add 1 vol Buffer C4 and 1 vol ethanol
8. Vortex for 30 s
9. Pipette 700 µL mixture onto a NucleoSpin Food Column placed in a Collection Tube
10. Centrifuge for 1 min at 11,000 x *g*
11. Discard the flow-through
12. Repeat steps 9-11
13. Add 400 µL Buffer CQW onto the spin column
14. Centrifuge for 1 min at 11,000 x g
15. Discard the flow-through
16. Add 700 µL Buffer C5 onto the spin column
17. Centrifuge for 1 min at 11,000 x g
18. Discard the flow-through
19. Add 200 µL Buffer C5 onto the spin column
20. Centrifuge for 2 min at 11,000 x g
21. Place the spin column into a new 1.5 mL microcentrifuge tube
22. Add 100 µL Elution Buffer CE (pre-heated to 70 °C) onto the membrane
23. Incubate for 5 min at room temperature (18-25 °C)
24. Centrifuge for 1 min a 11,000 x g
25. Proceed with amplification or keep frozen until further processing

2.3.2 Amplification

This protocol utilises a dual-indexing strategy (Kozich et al. 2013) amplifying the ITS2 region, using the primers ITS-S2F (Chen et al. 2010) and ITS4R (White et al. 1990). The primer sequences are as follows: forward: 5′-AATGATACGGCGACCACCGAGATCTACAC XXXXXXXX CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT-3′; reverse: 5′-CAAGCAGAAGACGGCATACGAGAT XXXXXXXX AGTCAGTCAG CC TCCTCCGCTTATTGATATGC-3′, where XXXXXX indicates the variable index sequences (Table 1). The detailed protocol is described below:

1. Sample index combinations should be planned beforehand according to the scheme in Figure 1
2. Prepare 3 x 10 µL reaction mixes for each sample containing (also see PCR sample design 2.3.2.1 below for details):
    - 5 µL 2 x Phusion Master Mix (New England Biolabs, Ipswich, MA, USA) or equivalent
    - 0.33 µM each of the forward and reverse primers (sample-specific combinations of forward and reverse index sequences)
    - 3.34 µL PCR grade water
    - 1 µL DNA template
3. Carry out the PCR with a programme of:
    - 95 °C for 4 min., then
    - 37 cycles of 95 °C for 40 sec.;
    - 49 °C for 40 sec.;
    - 72 °C for 40 sec. and
    - a final extension at 72°C for 5 min.
4. Combine the triplicate PCR reactions of each sample and mix well.

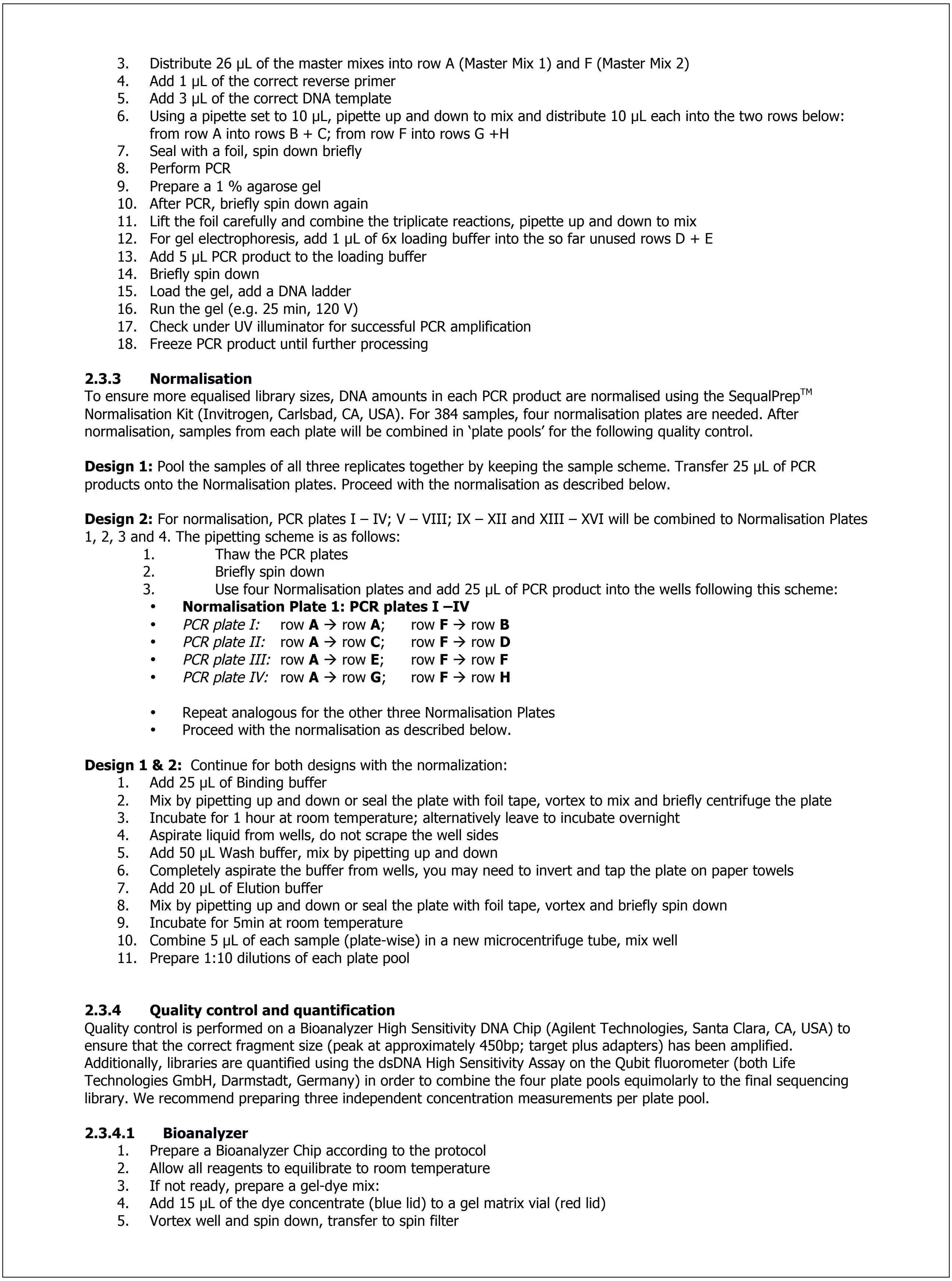
For quality control purposes, successful amplification can be checked on a 1 % agarose gel using 5 µL of the combined PCR product.

2.3.2.1 96-well PCR sample design

Design 1: Well-equipped laboratories with pipetting robots or 96-channel pipettes can directly fill each well with a different sample and generate three replicates of these. This will result in 4 x 3 replicate 96-well plates according to Figure 1 used for amplification. After amplification one can proceed with 2.3.3. Normalisation.

Design 2: For laboratories with little equipment for automated pipetting, the workflow described above is impractical, since manual pipetting in that format is time-intensive and pipetting errors can be easily introduced. To facilitate the process, we recommend to work with all triplicates but only 24 samples on one 96 well plate (Figure 2). This way, 16 PCR plates will be produced, but pipetting effort is minimized. PCR plate labelling is therefore of utter importance, for example with roman numbers, I – XVI to be able to map the samples back to the scheme in Figure 1. The complete workflow is shown schematically in Figure 2 and described in the following:

1. Prepare two PCR master mixes, each containing one forward primer, corresponding to the samples you want to amplify; each master mix contains:
    - 200 µL 2 x Phusion Master Mix (New England Biolabs, Ipswich, MA, USA) or equivalent
    - 13.2 µL forward primer
    - 133.6 µL PCR grade water
2. Place a new PCR plate into a cooling block



3. Distribute 26 µL of the master mixes into row A (Master Mix 1) and F (Master Mix 2)
4. Add 1 µL of the correct reverse primer
5. Add 3 µL of the correct DNA template
6. Using a pipette set to 10 µL, pipette up and down to mix and distribute 10 µL each into the two rows below: from row A into rows B + C; from row F into rows G +H
7. Seal with a foil, spin down briefly
8. Perform PCR
9. Prepare a 1 % agarose gel
10. After PCR, briefly spin down again
11. Lift the foil carefully and combine the triplicate reactions, pipette up and down to mix
12. For gel electrophoresis, add 1 µL of 6x loading buffer into the so far unused rows D + E
13. Add 5 µL PCR product to the loading buffer
14. Briefly spin down
15. Load the gel, add a DNA ladder
16. Run the gel (e.g. 25 min, 120 V)
17. Check under UV illuminator for successful PCR amplification
18. Freeze PCR product until further processing

2.3.3 Normalisation

To ensure more equalised library sizes, DNA amounts in each PCR product are normalised using the SequalPrep^{TM} Normalisation Kit (Invitrogen, Carlsbad, CA, USA). For 384 samples, four normalisation plates are needed. After normalisation, samples from each plate will be combined in 'plate pools' for the following quality control.

Design 1: Pool the samples of all three replicates together by keeping the sample scheme. Transfer 25 µL of PCR products onto the Normalisation plates. Proceed with the normalisation as described below.

Design 2: For normalisation, PCR plates I – IV; V – VIII; IX – XII and XIII – XVI will be combined to Normalisation Plates 1, 2, 3 and 4. The pipetting scheme is as follows:

1. Thaw the PCR plates
2. Briefly spin down
3. Use four Normalisation plates and add 25 µL of PCR product into the wells following this scheme:

- **Normalisation Plate 1: PCR plates I –IV**
- *PCR plate I:* row **A** → row **A**; row **F** → row **B**
- *PCR plate II:* row **A** → row **C**; row **F** → row **D**
- *PCR plate III:* row **A** → row **E**; row **F** → row **F**
- *PCR plate IV:* row **A** → row **G**; row **F** → row **H**

- Repeat analogous for the other three Normalisation Plates
- Proceed with the normalisation as described below.

Design 1 & 2: Continue for both designs with the normalization:

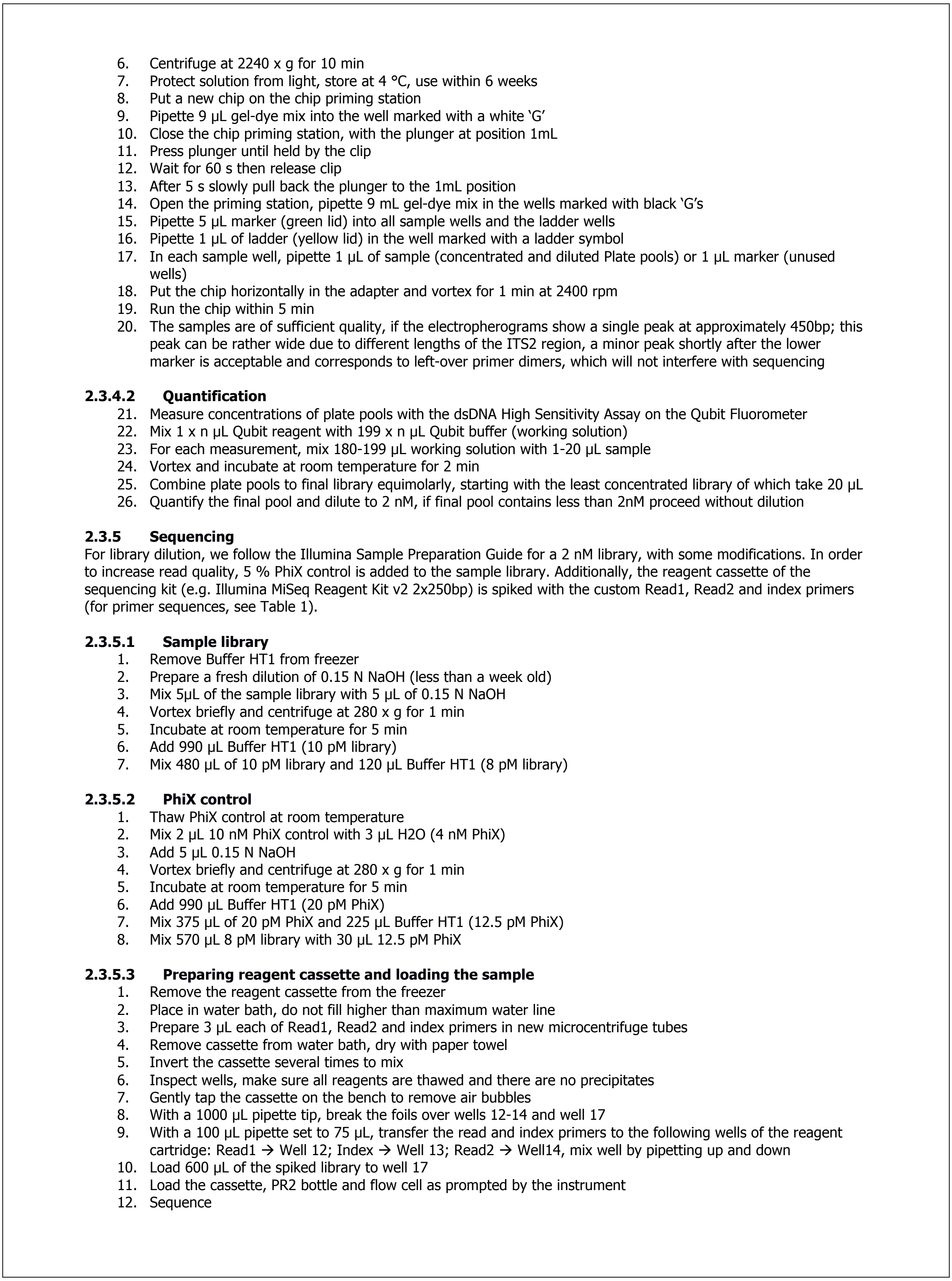
1. Add 25 µL of Binding buffer
2. Mix by pipetting up and down or seal the plate with foil tape, vortex to mix and briefly centrifuge the plate
3. Incubate for 1 hour at room temperature; alternatively leave to incubate overnight
4. Aspirate liquid from wells, do not scrape the well sides
5. Add 50 µL Wash buffer, mix by pipetting up and down
6. Completely aspirate the buffer from wells, you may need to invert and tap the plate on paper towels
7. Add 20 µL of Elution buffer
8. Mix by pipetting up and down or seal the plate with foil tape, vortex and briefly spin down
9. Incubate for 5min at room temperature
10. Combine 5 µL of each sample (plate-wise) in a new microcentrifuge tube, mix well
11. Prepare 1:10 dilutions of each plate pool

2.3.4 Quality control and quantification

Quality control is performed on a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA, USA) to ensure that the correct fragment size (peak at approximately 450bp; target plus adapters) has been amplified. Additionally, libraries are quantified using the dsDNA High Sensitivity Assay on the Qubit fluorometer (both Life Technologies GmbH, Darmstadt, Germany) in order to combine the four plate pools equimolarly to the final sequencing library. We recommend preparing three independent concentration measurements per plate pool.

2.3.4.1 Bioanalyzer

1. Prepare a Bioanalyzer Chip according to the protocol
2. Allow all reagents to equilibrate to room temperature
3. If not ready, prepare a gel-dye mix:
4. Add 15 µL of the dye concentrate (blue lid) to a gel matrix vial (red lid)
5. Vortex well and spin down, transfer to spin filter



6. Centrifuge at 2240 x g for 10 min
7. Protect solution from light, store at 4 °C, use within 6 weeks
8. Put a new chip on the chip priming station
9. Pipette 9 µL gel-dye mix into the well marked with a white 'G'
10. Close the chip priming station, with the plunger at position 1mL
11. Press plunger until held by the clip
12. Wait for 60 s then release clip
13. After 5 s slowly pull back the plunger to the 1mL position
14. Open the priming station, pipette 9 mL gel-dye mix in the wells marked with black 'G's
15. Pipette 5 µL marker (green lid) into all sample wells and the ladder wells
16. Pipette 1 µL of ladder (yellow lid) in the well marked with a ladder symbol
17. In each sample well, pipette 1 µL of sample (concentrated and diluted Plate pools) or 1 µL marker (unused wells)
18. Put the chip horizontally in the adapter and vortex for 1 min at 2400 rpm
19. Run the chip within 5 min
20. The samples are of sufficient quality, if the electropherograms show a single peak at approximately 450bp; this peak can be rather wide due to different lengths of the ITS2 region, a minor peak shortly after the lower marker is acceptable and corresponds to left-over primer dimers, which will not interfere with sequencing

2.3.4.2 Quantification

21. Measure concentrations of plate pools with the dsDNA High Sensitivity Assay on the Qubit Fluorometer
22. Mix 1 x n µL Qubit reagent with 199 x n µL Qubit buffer (working solution)
23. For each measurement, mix 180-199 µL working solution with 1-20 µL sample
24. Vortex and incubate at room temperature for 2 min
25. Combine plate pools to final library equimolarly, starting with the least concentrated library of which take 20 µL
26. Quantify the final pool and dilute to 2 nM, if final pool contains less than 2nM proceed without dilution

2.3.5 Sequencing

For library dilution, we follow the Illumina Sample Preparation Guide for a 2 nM library, with some modifications. In order to increase read quality, 5 % PhiX control is added to the sample library. Additionally, the reagent cassette of the sequencing kit (e.g. Illumina MiSeq Reagent Kit v2 2x250bp) is spiked with the custom Read1, Read2 and index primers (for primer sequences, see Table 1).

2.3.5.1 Sample library

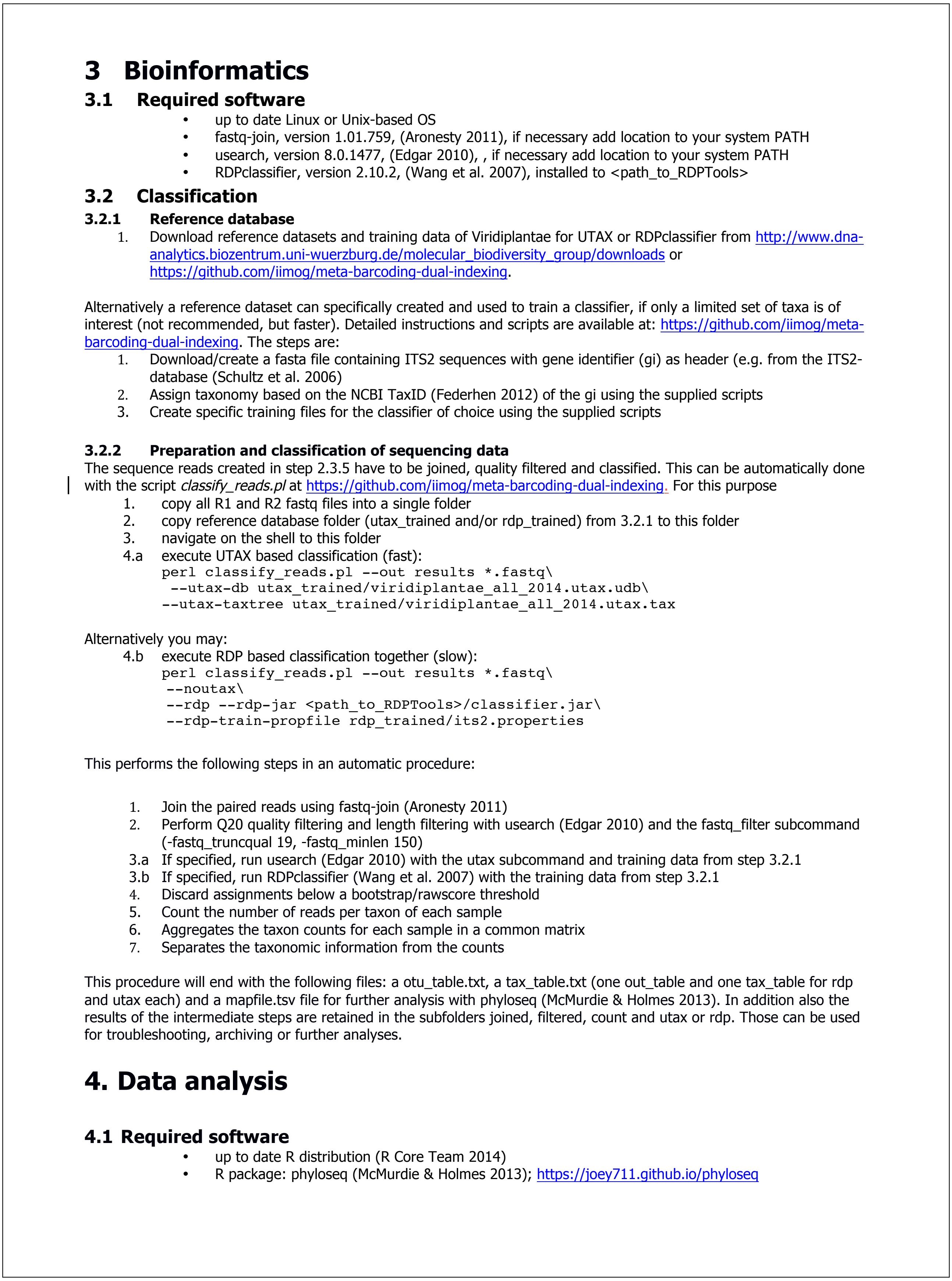
1. Remove Buffer HT1 from freezer
2. Prepare a fresh dilution of 0.15 N NaOH (less than a week old)
3. Mix 5µL of the sample library with 5 µL of 0.15 N NaOH
4. Vortex briefly and centrifuge at 280 x g for 1 min
5. Incubate at room temperature for 5 min
6. Add 990 µL Buffer HT1 (10 pM library)
7. Mix 480 µL of 10 pM library and 120 µL Buffer HT1 (8 pM library)

2.3.5.2 PhiX control

1. Thaw PhiX control at room temperature
2. Mix 2 µL 10 nM PhiX control with 3 µL H2O (4 nM PhiX)
3. Add 5 µL 0.15 N NaOH
4. Vortex briefly and centrifuge at 280 x g for 1 min
5. Incubate at room temperature for 5 min
6. Add 990 µL Buffer HT1 (20 pM PhiX)
7. Mix 375 µL of 20 pM PhiX and 225 µL Buffer HT1 (12.5 pM PhiX)
8. Mix 570 µL 8 pM library with 30 µL 12.5 pM PhiX

2.3.5.3 Preparing reagent cassette and loading the sample

1. Remove the reagent cassette from the freezer
2. Place in water bath, do not fill higher than maximum water line
3. Prepare 3 µL each of Read1, Read2 and index primers in new microcentrifuge tubes
4. Remove cassette from water bath, dry with paper towel
5. Invert the cassette several times to mix
6. Inspect wells, make sure all reagents are thawed and there are no precipitates
7. Gently tap the cassette on the bench to remove air bubbles
8. With a 1000 µL pipette tip, break the foils over wells 12-14 and well 17
9. With a 100 µL pipette set to 75 µL, transfer the read and index primers to the following wells of the reagent cartridge: Read1 → Well 12; Index → Well 13; Read2 → Well14, mix well by pipetting up and down
10. Load 600 µL of the spiked library to well 17
11. Load the cassette, PR2 bottle and flow cell as prompted by the instrument
12. Sequence



3 Bioinformatics

3.1 Required software

- up to date Linux or Unix-based OS
- fastq-join, version 1.01.759, (Aronesty 2011), if necessary add location to your system PATH
- usearch, version 8.0.1477, (Edgar 2010), , if necessary add location to your system PATH
- RDPclassifier, version 2.10.2, (Wang et al. 2007), installed to <path_to_RDPTools>

3.2 Classification

3.2.1 Reference database

1. Download reference datasets and training data of Viridiplantae for UTAX or RDPclassifier from http://www.dna-analytics.biozentrum.uni-wuerzburg.de/molecular_biodiversity_group/downloads or https://github.com/iimog/meta-barcoding-dual-indexing.

Alternatively a reference dataset can specifically created and used to train a classifier, if only a limited set of taxa is of interest (not recommended, but faster). Detailed instructions and scripts are available at: https://github.com/iimog/meta-barcoding-dual-indexing. The steps are:

1. Download/create a fasta file containing ITS2 sequences with gene identifier (gi) as header (e.g. from the ITS2-database (Schultz et al. 2006)
2. Assign taxonomy based on the NCBI TaxID (Federhen 2012) of the gi using the supplied scripts
3. Create specific training files for the classifier of choice using the supplied scripts

3.2.2 Preparation and classification of sequencing data

The sequence reads created in step 2.3.5 have to be joined, quality filtered and classified. This can be automatically done with the script *classify_reads.pl* at https://github.com/iimog/meta-barcoding-dual-indexing. For this purpose

1. copy all R1 and R2 fastq files into a single folder
2. copy reference database folder (utax_trained and/or rdp_trained) from 3.2.1 to this folder
3. navigate on the shell to this folder

4.a execute UTAX based classification (fast):

```
perl classify_reads.pl --out results *.fastq\
 --utax-db utax_trained/viridiplantae_all_2014.utax.udb\
--utax-taxtree utax_trained/viridiplantae_all_2014.utax.tax
```

Alternatively you may:

4.b execute RDP based classification together (slow):

```
perl classify_reads.pl --out results *.fastq\
 --noutax\
 --rdp --rdp-jar <path_to_RDPTools>/classifier.jar\
 --rdp-train-propfile rdp_trained/its2.properties
```

This performs the following steps in an automatic procedure:

1. Join the paired reads using fastq-join (Aronesty 2011)
2. Perform Q20 quality filtering and length filtering with usearch (Edgar 2010) and the fastq_filter subcommand (-fastq_truncqual 19, -fastq_minlen 150)

3.a If specified, run usearch (Edgar 2010) with the utax subcommand and training data from step 3.2.1

3.b If specified, run RDPclassifier (Wang et al. 2007) with the training data from step 3.2.1

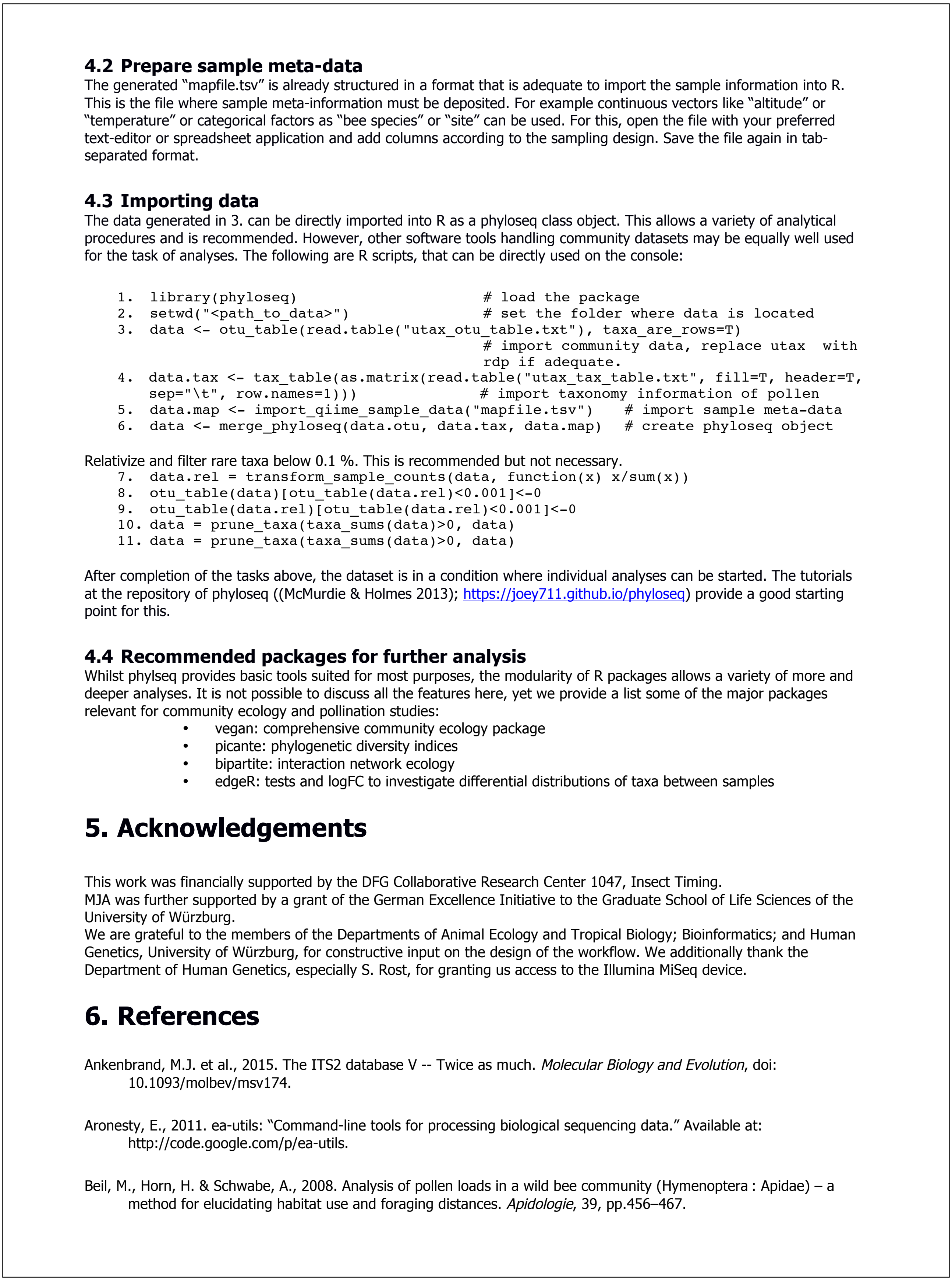
4. Discard assignments below a bootstrap/rawscore threshold
5. Count the number of reads per taxon of each sample
6. Aggregates the taxon counts for each sample in a common matrix
7. Separates the taxonomic information from the counts

This procedure will end with the following files: a otu_table.txt, a tax_table.txt (one out_table and one tax_table for rdp and utax each) and a mapfile.tsv file for further analysis with phyloseq (McMurdie & Holmes 2013). In addition also the results of the intermediate steps are retained in the subfolders joined, filtered, count and utax or rdp. Those can be used for troubleshooting, archiving or further analyses.

4. Data analysis

4.1 Required software

- up to date R distribution (R Core Team 2014)
- R package: phyloseq (McMurdie & Holmes 2013); https://joey711.github.io/phyloseq



4.2 Prepare sample meta-data

The generated "mapfile.tsv" is already structured in a format that is adequate to import the sample information into R. This is the file where sample meta-information must be deposited. For example continuous vectors like "altitude" or "temperature" or categorical factors as "bee species" or "site" can be used. For this, open the file with your preferred text-editor or spreadsheet application and add columns according to the sampling design. Save the file again in tab-separated format.

4.3 Importing data

The data generated in 3. can be directly imported into R as a phyloseq class object. This allows a variety of analytical procedures and is recommended. However, other software tools handling community datasets may be equally well used for the task of analyses. The following are R scripts, that can be directly used on the console:

```
1.  library(phyloseq)                     # load the package
2.  setwd("<path_to_data>")               # set the folder where data is located
3.  data <- otu_table(read.table("utax_otu_table.txt"), taxa_are_rows=T)
                                          # import community data, replace utax  with
                                          rdp if adequate.
4.  data.tax <- tax_table(as.matrix(read.table("utax_tax_table.txt", fill=T, header=T,
    sep="\t", row.names=1)))              # import taxonomy information of pollen
5.  data.map <- import_qiime_sample_data("mapfile.tsv")    # import sample meta-data
6.  data <- merge_phyloseq(data.otu, data.tax, data.map)   # create phyloseq object
```

Relativize and filter rare taxa below 0.1 %. This is recommended but not necessary.

```
7.  data.rel = transform_sample_counts(data, function(x) x/sum(x))
8.  otu_table(data)[otu_table(data.rel)<0.001]<-0
9.  otu_table(data.rel)[otu_table(data.rel)<0.001]<-0
10. data = prune_taxa(taxa_sums(data)>0, data)
11. data = prune_taxa(taxa_sums(data)>0, data)
```

After completion of the tasks above, the dataset is in a condition where individual analyses can be started. The tutorials at the repository of phyloseq ((McMurdie & Holmes 2013); https://joey711.github.io/phyloseq) provide a good starting point for this.

4.4 Recommended packages for further analysis

Whilst phylseq provides basic tools suited for most purposes, the modularity of R packages allows a variety of more and deeper analyses. It is not possible to discuss all the features here, yet we provide a list some of the major packages relevant for community ecology and pollination studies:

- vegan: comprehensive community ecology package
- picante: phylogenetic diversity indices
- bipartite: interaction network ecology
- edgeR: tests and logFC to investigate differential distributions of taxa between samples

5. Acknowledgements

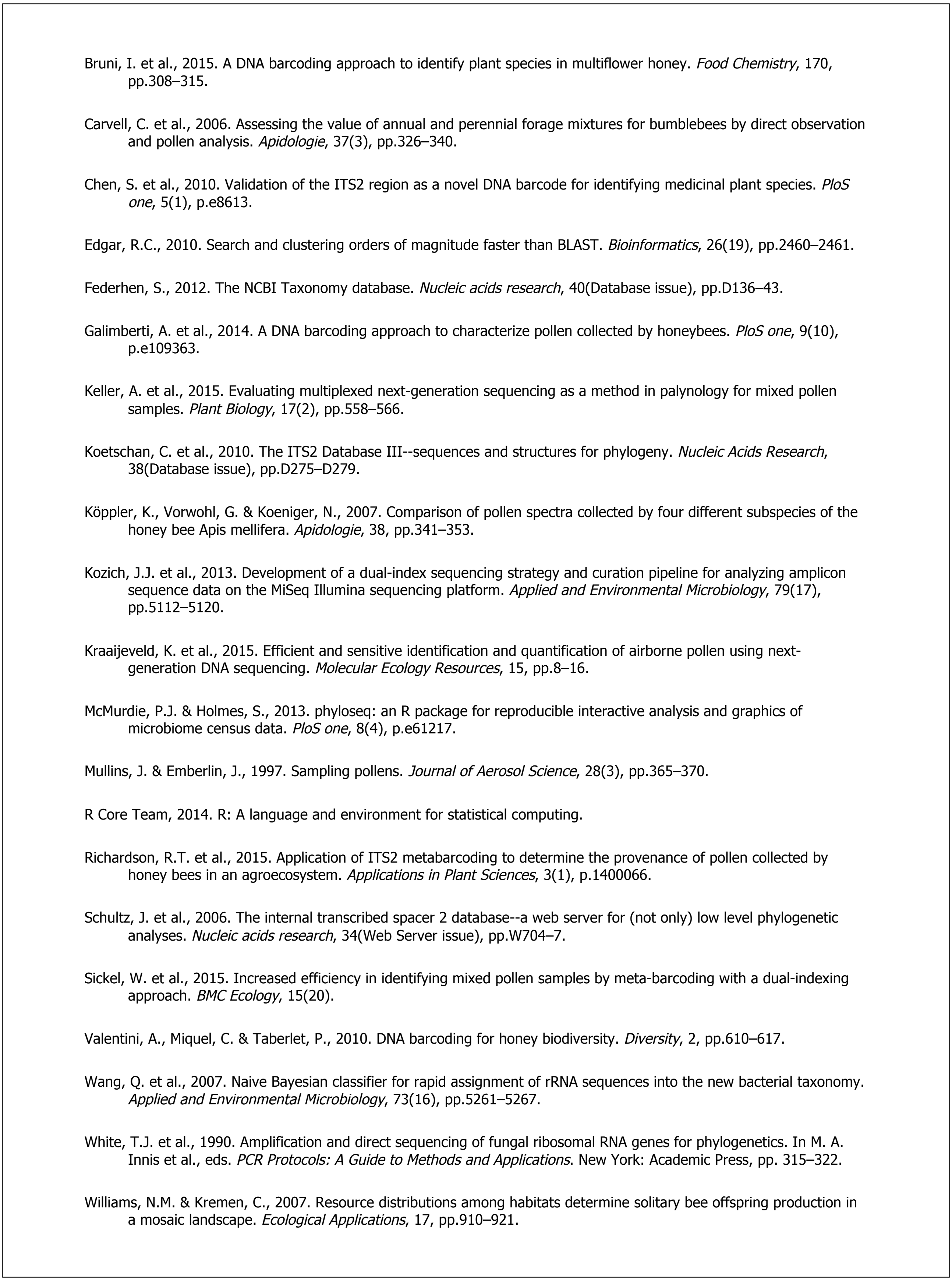
This work was financially supported by the DFG Collaborative Research Center 1047, Insect Timing.
MJA was further supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences of the University of Würzburg.
We are grateful to the members of the Departments of Animal Ecology and Tropical Biology; Bioinformatics; and Human Genetics, University of Würzburg, for constructive input on the design of the workflow. We additionally thank the Department of Human Genetics, especially S. Rost, for granting us access to the Illumina MiSeq device.

6. References

Ankenbrand, M.J. et al., 2015. The ITS2 database V -- Twice as much. *Molecular Biology and Evolution*, doi: 10.1093/molbev/msv174.

Aronesty, E., 2011. ea-utils: "Command-line tools for processing biological sequencing data." Available at: http://code.google.com/p/ea-utils.

Beil, M., Horn, H. & Schwabe, A., 2008. Analysis of pollen loads in a wild bee community (Hymenoptera : Apidae) – a method for elucidating habitat use and foraging distances. *Apidologie*, 39, pp.456–467.



Bruni, I. et al., 2015. A DNA barcoding approach to identify plant species in multiflower honey. *Food Chemistry*, 170, pp.308–315.

Carvell, C. et al., 2006. Assessing the value of annual and perennial forage mixtures for bumblebees by direct observation and pollen analysis. *Apidologie*, 37(3), pp.326–340.

Chen, S. et al., 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS one*, 5(1), p.e8613.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), pp.2460–2461.

Federhen, S., 2012. The NCBI Taxonomy database. *Nucleic acids research*, 40(Database issue), pp.D136–43.

Galimberti, A. et al., 2014. A DNA barcoding approach to characterize pollen collected by honeybees. *PloS one*, 9(10), p.e109363.

Keller, A. et al., 2015. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology*, 17(2), pp.558–566.

Koetschan, C. et al., 2010. The ITS2 Database III--sequences and structures for phylogeny. *Nucleic Acids Research*, 38(Database issue), pp.D275–D279.

Köppler, K., Vorwohl, G. & Koeniger, N., 2007. Comparison of pollen spectra collected by four different subspecies of the honey bee Apis mellifera. *Apidologie*, 38, pp.341–353.

Kozich, J.J. et al., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*, 79(17), pp.5112–5120.

Kraaijeveld, K. et al., 2015. Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resources*, 15, pp.8–16.

McMurdie, P.J. & Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4), p.e61217.

Mullins, J. & Emberlin, J., 1997. Sampling pollens. *Journal of Aerosol Science*, 28(3), pp.365–370.

R Core Team, 2014. R: A language and environment for statistical computing.

Richardson, R.T. et al., 2015. Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Applications in Plant Sciences*, 3(1), p.1400066.

Schultz, J. et al., 2006. The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. *Nucleic acids research*, 34(Web Server issue), pp.W704–7.

Sickel, W. et al., 2015. Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology*, 15(20).

Valentini, A., Miquel, C. & Taberlet, P., 2010. DNA barcoding for honey biodiversity. *Diversity*, 2, pp.610–617.

Wang, Q. et al., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), pp.5261–5267.

White, T.J. et al., 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. A. Innis et al., eds. *PCR Protocols: A Guide to Methods and Applications*. New York: Academic Press, pp. 315–322.

Williams, N.M. & Kremen, C., 2007. Resource distributions among habitats determine solitary bee offspring production in a mosaic landscape. *Ecological Applications*, 17, pp.910–921.

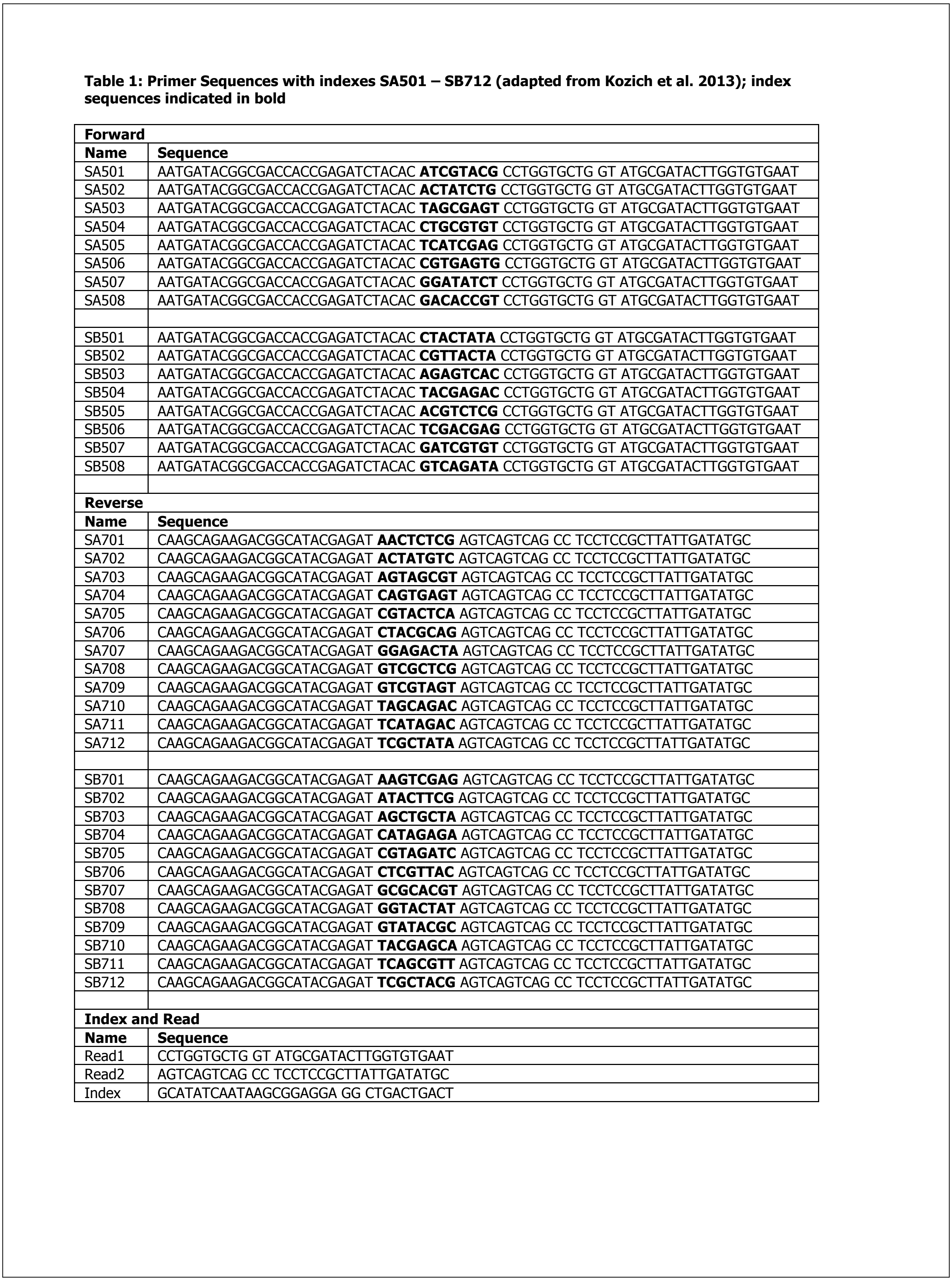


Table 1: Primer Sequences with indexes SA501 – SB712 (adapted from Kozich et al. 2013); index sequences indicated in bold

Forward	
Name	**Sequence**
SA501	AATGATACGGCGACCACCGAGATCTACAC **ATCGTACG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA502	AATGATACGGCGACCACCGAGATCTACAC **ACTATCTG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA503	AATGATACGGCGACCACCGAGATCTACAC **TAGCGAGT** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA504	AATGATACGGCGACCACCGAGATCTACAC **CTGCGTGT** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA505	AATGATACGGCGACCACCGAGATCTACAC **TCATCGAG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA506	AATGATACGGCGACCACCGAGATCTACAC **CGTGAGTG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA507	AATGATACGGCGACCACCGAGATCTACAC **GGATATCT** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SA508	AATGATACGGCGACCACCGAGATCTACAC **GACACCGT** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB501	AATGATACGGCGACCACCGAGATCTACAC **CTACTATA** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB502	AATGATACGGCGACCACCGAGATCTACAC **CGTTACTA** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB503	AATGATACGGCGACCACCGAGATCTACAC **AGAGTCAC** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB504	AATGATACGGCGACCACCGAGATCTACAC **TACGAGAC** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB505	AATGATACGGCGACCACCGAGATCTACAC **ACGTCTCG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB506	AATGATACGGCGACCACCGAGATCTACAC **TCGACGAG** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB507	AATGATACGGCGACCACCGAGATCTACAC **GATCGTGT** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
SB508	AATGATACGGCGACCACCGAGATCTACAC **GTCAGATA** CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
Reverse	
Name	**Sequence**
SA701	CAAGCAGAAGACGGCATACGAGAT **AACTCTCG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA702	CAAGCAGAAGACGGCATACGAGAT **ACTATGTC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA703	CAAGCAGAAGACGGCATACGAGAT **AGTAGCGT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA704	CAAGCAGAAGACGGCATACGAGAT **CAGTGAGT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA705	CAAGCAGAAGACGGCATACGAGAT **CGTACTCA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA706	CAAGCAGAAGACGGCATACGAGAT **CTACGCAG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA707	CAAGCAGAAGACGGCATACGAGAT **GGAGACTA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA708	CAAGCAGAAGACGGCATACGAGAT **GTCGCTCG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA709	CAAGCAGAAGACGGCATACGAGAT **GTCGTAGT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA710	CAAGCAGAAGACGGCATACGAGAT **TAGCAGAC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA711	CAAGCAGAAGACGGCATACGAGAT **TCATAGAC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SA712	CAAGCAGAAGACGGCATACGAGAT **TCGCTATA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB701	CAAGCAGAAGACGGCATACGAGAT **AAGTCGAG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB702	CAAGCAGAAGACGGCATACGAGAT **ATACTTCG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB703	CAAGCAGAAGACGGCATACGAGAT **AGCTGCTA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB704	CAAGCAGAAGACGGCATACGAGAT **CATAGAGA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB705	CAAGCAGAAGACGGCATACGAGAT **CGTAGATC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB706	CAAGCAGAAGACGGCATACGAGAT **CTCGTTAC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB707	CAAGCAGAAGACGGCATACGAGAT **GCGCACGT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB708	CAAGCAGAAGACGGCATACGAGAT **GGTACTAT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB709	CAAGCAGAAGACGGCATACGAGAT **GTATACGC** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB710	CAAGCAGAAGACGGCATACGAGAT **TACGAGCA** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB711	CAAGCAGAAGACGGCATACGAGAT **TCAGCGTT** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
SB712	CAAGCAGAAGACGGCATACGAGAT **TCGCTACG** AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
Index and Read	
Name	**Sequence**
Read1	CCTGGTGCTG GT ATGCGATACTTGGTGTGAAT
Read2	AGTCAGTCAG CC TCCTCCGCTTATTGATATGC
Index	GCATATCAATAAGCGGAGGA GG CTGACTGACT

Figure 1 Planning scheme for samples and the corresponding index-combinations. Roman numbers indicate PCR plate numbers, bold Arabian numbers on 96 well plates indicate Normalisation plate number.

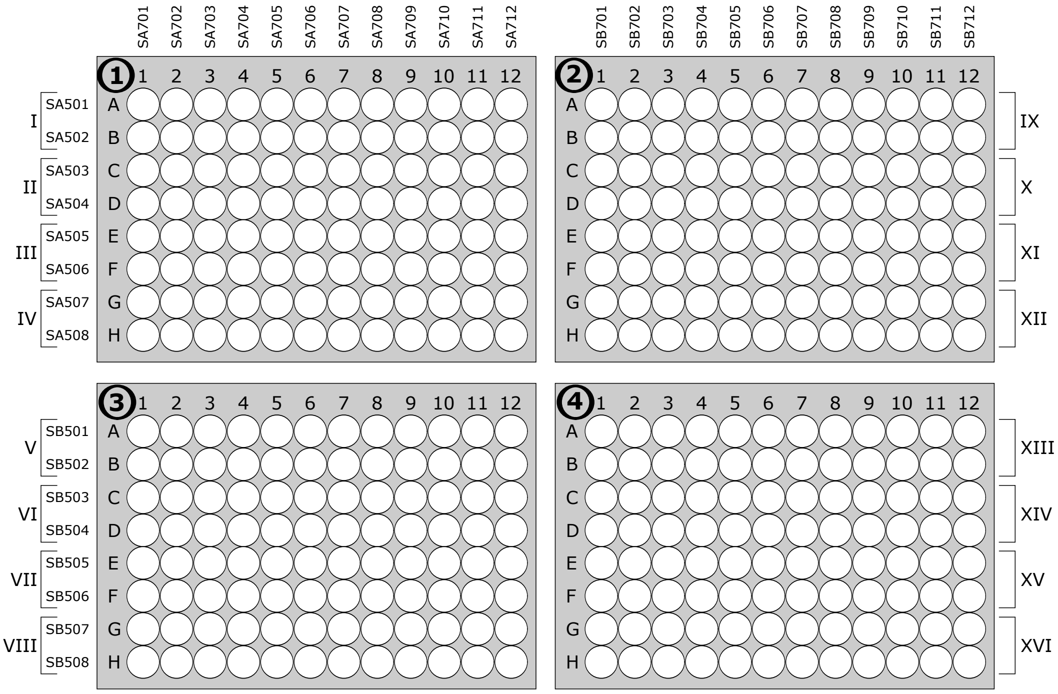
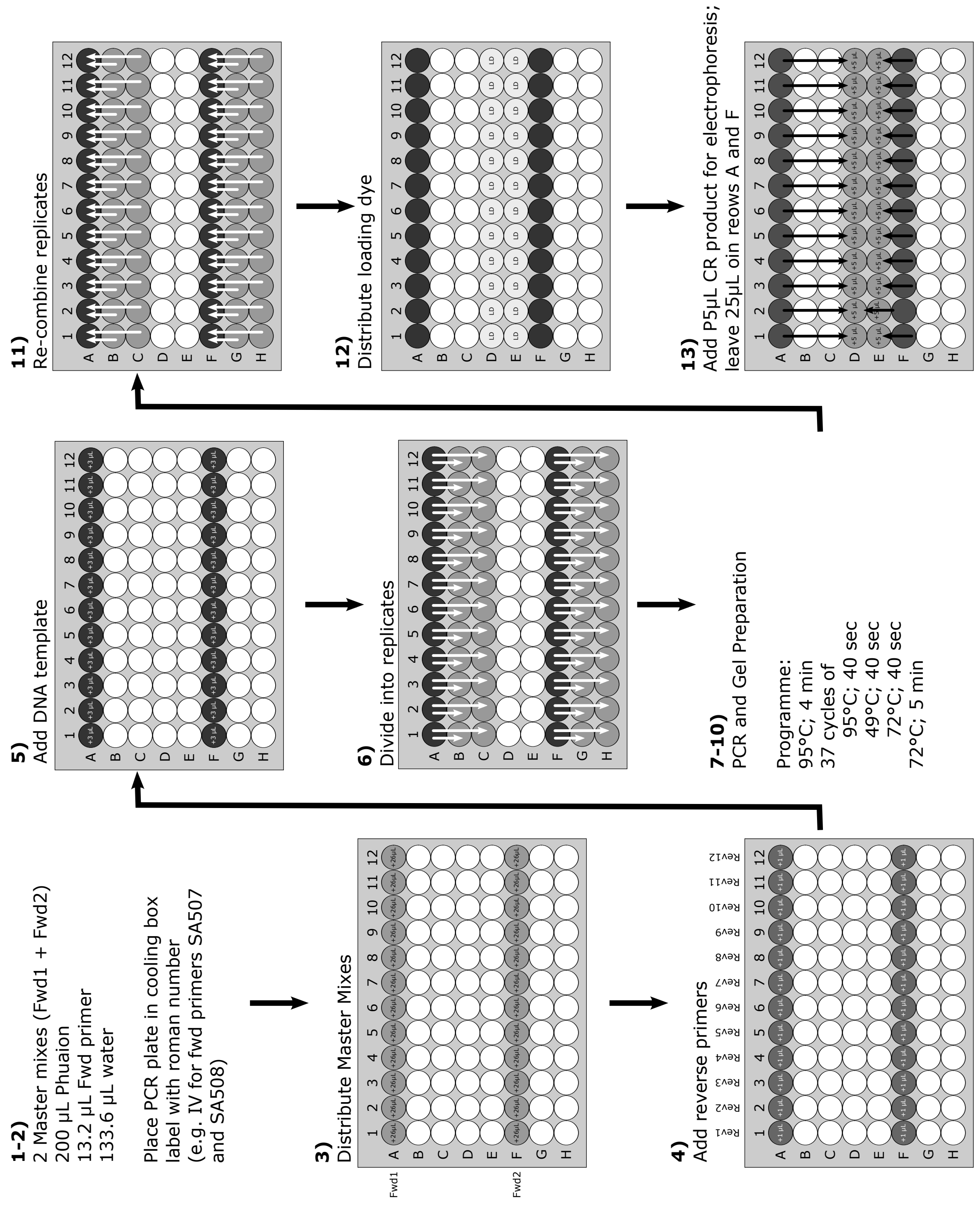


Figure 2 Detailed workflow (schematic), suitable for laboratories with limited access to equipment for automated pipetting. Bold numbers indicated step number of Design 2 in sub-chapter 2.3.2.1



5.5 DNA-METABARCODING - EIN NEUER BLICK AUF ORGANISMISCHE DIVERSITÄT

– published in *BIOspektrum* –

Permission for legal second publication has been granted by the publisher with License Numbers 4092970266068 and 4092970135081. This publication is also part of Sickel (2017). It is part of both theses because of complementary contributions.

Genetische Ökologie

DNA-Metabarcoding – ein neuer Blick auf organismische Diversität

ALEXANDER KELLER, GUDRUN GRIMMER, WIEBKE SICKEL,
MARKUS J. ANKENBRAND
LEHRSTUHL FÜR TIERÖKOLOGIE UND TROPENBIOLOGIE, BIOZENTRUM,
UNIVERSITÄT WÜRZBURG

Taxon identification is one of the fundamental challenges in biological research. Usually, classifications are based on specimen morphology, sometimes supported by their behaviour, ecology or biochemistry. Technological advances now allow using genomic fragments as a taxon barcode. With the latest developments of high-throughput sequencers this can go even further: identifying complete assemblages simultaneously, with various applications in ecology, conservation, forensics and health security.

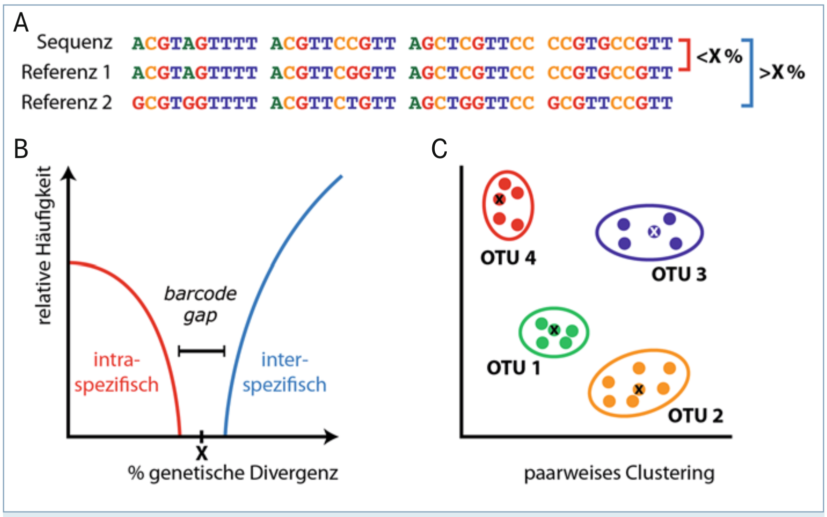
DOI: 10.1007/s12268-016-0669-0

■ Die Identifikation von Organismen stellt eine der grundlegendsten und ältesten Herausforderungen in der biologischen Forschung dar. Traditionell wird diese Erkennung und Abgrenzung von anderen Lebewesen über morphologische Merkmale durchgeführt, ggf. werden je nach taxonomischer Gruppe auch ethologische, biochemische oder ökologische Informationen zurate gezogen. Durch die technologischen Entwicklungen in den vergangenen Jahren stehen uns heute zusätzlich genomische Daten in Form von DNA-Sequenzen zur Verfügung, die auch bei der Klassifizierung und Unterscheidung von Organismen hilfreich sein können.

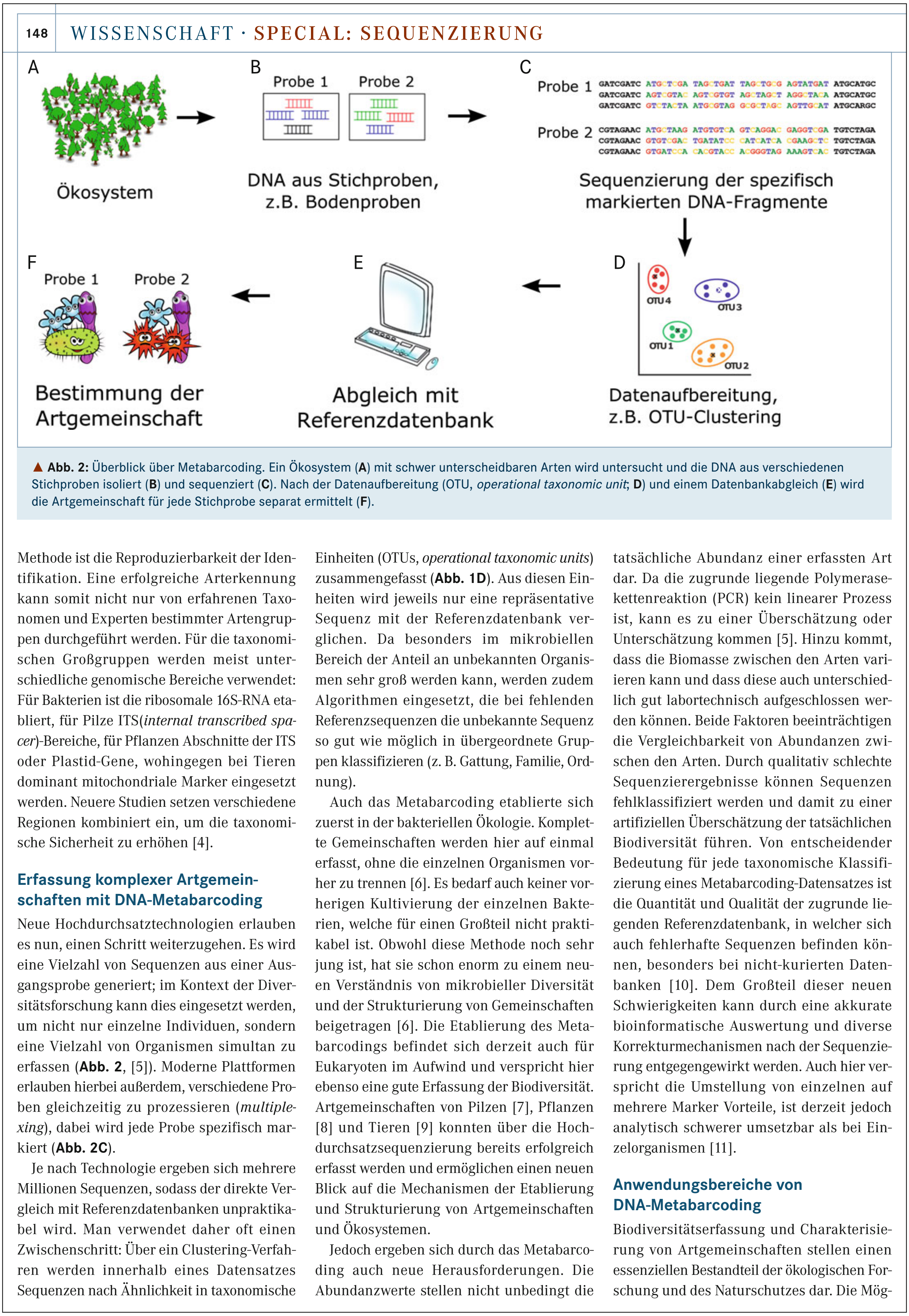
DNA-Barcoding unterstützt die traditionelle Arterkennung

In der Diversitätsforschung wurde die Sequenzierung genomischer DNA-Fragmente schon relativ früh eingesetzt, um die evolutive Geschichte von Organismen zu rekonstruieren [1]. Dabei werden Sequenzen unterschiedlicher Organismen miteinander verglichen, Unterschiede ermittelt und diese zur Erstellung eines phylogenetischen Stammbaums verwendet. Vor allem in der Mikrobiologie etablierte sich diese Methode schnell, da sie nicht mehr auf die wenigen erfassbaren Merkmale der Individuen angewiesen war [1]. Dementsprechend verwundert es nicht, dass erste Schritte zur Katalogisierung von Organismen anhand von Sequenzen auch in mikrobiologischen Werken zu finden sind [2]. Erst im Jahr 2003 wurde diese Methode unter dem Namen DNA-Barcoding auch für höhere Eukaryoten etabliert [3]. Inzwischen ist die Methode weit verbreitet und wird durch zahlreiche Initiativen gestützt. Die grundlegenden Ziele des DNA-Barcodings sind die flächendeckende Katalogisierung der organismischen Diversität und deren Nutzung als Referenz für weiterführende Fragestellungen.

Das Prinzip des DNA-Barcodings besteht darin, ein kurzes Fragment der genomischen DNA zu analysieren, das repräsentativ für eine bestimmte Art ist und eindeutig auf diese zurückgeführt werden kann. Über einen bioinformatischen Vergleich mittels eines Schwellenwertes (*barcoding gap*) kann die Identität einer unbekannten Sequenz anhand einer Referenzdatenbank bestimmt werden (**Abb. 1A**). Dieser Schwellenwert wird so definiert, dass intraspezifische von interspezifischer genomischer Variation unterschieden wird (**Abb. 1B**). Ein großer Vorteil dieser


A
Sequenz ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
Referenz 1 ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
Referenz 2 GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
<X %
>X %
B
relative Häufigkeit
barcode gap
intra-spezifisch
inter-spezifisch
X
% genetische Divergenz
C
OTU 4
OTU 3
OTU 1
OTU 2
paarweises Clustering

▲ **Abb. 1:** Bioinformatischer Ablauf einer DNA-Barcoding-Studie. **A**, Sequenzidentitäten mit Referenzen kleiner dem Schwellenwert X gelten als erfolgreiche Artidentifizierung. **B**, X wird bestimmt durch die *barcode gap* zwischen der Variation innerhalb einer Art und zu anderen Arten. **C**, Einordnung ähnlicher Sequenzen in taxonomische Einheiten (OTU, *operational taxonomic unit*) eines Metabarcoding-Datensatzes; nur eine repräsentative Sequenz wird mit der Datenbank abgeglichen.



A
Ökosystem
B
Probe 1
Probe 2
DNA aus Stichproben, z.B. Bodenproben
C
Probe 1
GATCGATC ATGCTCGA TAGCTGAT TAGCTGCG AGTATGAT ATGCATGC
GATCGATC AGTCGTAC AGTCGTGT AGCTAGCT AGGCTACA ATGCATGC
GATCGATC GTCTACTA ATGCGTAG GCGCTAGC AGTTGCAT ATGCARGC
Probe 2
CGTAGAAC ATGCTAAG ATGTGTCA GTCAGGAC GAGGTCGA TGTCTAGA
CGTAGAAC GTGTCGAC TGATATCC CATCATCA CGAAGCTC TGTCTAGA
CGTAGAAC GTGATCCA CACGTACC ACGGGTAG AAAGTCAC TGTCTAGA
Sequenzierung der spezifisch markierten DNA-Fragmente
D
OTU 4
OTU 3
OTU 1
OTU 2
Datenaufbereitung, z.B. OTU-Clustering
E
Abgleich mit Referenzdatenbank
F
Probe 1
Probe 2
Bestimmung der Artgemeinschaft

▲ **Abb. 2:** Überblick über Metabarcoding. Ein Ökosystem (**A**) mit schwer unterscheidbaren Arten wird untersucht und die DNA aus verschiedenen Stichproben isoliert (**B**) und sequenziert (**C**). Nach der Datenaufbereitung (OTU, *operational taxonomic unit*; **D**) und einem Datenbankabgleich (**E**) wird die Artgemeinschaft für jede Stichprobe separat ermittelt (**F**).

Methode ist die Reproduzierbarkeit der Identifikation. Eine erfolgreiche Arterkennung kann somit nicht nur von erfahrenen Taxonomen und Experten bestimmter Artengruppen durchgeführt werden. Für die taxonomischen Großgruppen werden meist unterschiedliche genomische Bereiche verwendet: Für Bakterien ist die ribosomale 16S-RNA etabliert, für Pilze ITS(*internal transcribed spacer*)-Bereiche, für Pflanzen Abschnitte der ITS oder Plastid-Gene, wohingegen bei Tieren dominant mitochondriale Marker eingesetzt werden. Neuere Studien setzen verschiedene Regionen kombiniert ein, um die taxonomische Sicherheit zu erhöhen [4].

Erfassung komplexer Artgemeinschaften mit DNA-Metabarcoding

Neue Hochdurchsatztechnologien erlauben es nun, einen Schritt weiterzugehen. Es wird eine Vielzahl von Sequenzen aus einer Ausgangsprobe generiert; im Kontext der Diversitätsforschung kann dies eingesetzt werden, um nicht nur einzelne Individuen, sondern eine Vielzahl von Organismen simultan zu erfassen (**Abb. 2**, [5]). Moderne Plattformen erlauben hierbei außerdem, verschiedene Proben gleichzeitig zu prozessieren (*multiplexing*), dabei wird jede Probe spezifisch markiert (**Abb. 2C**).

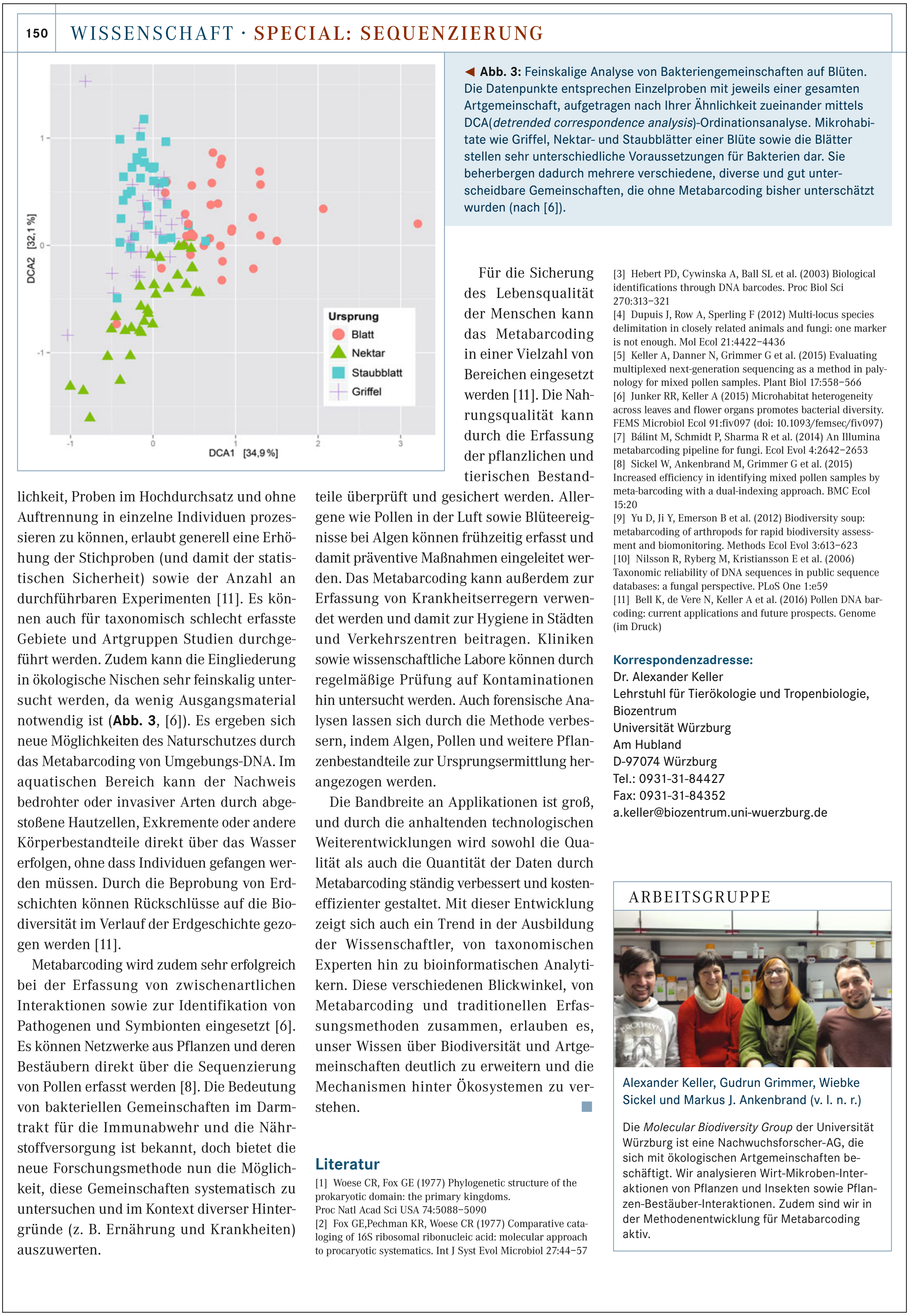
Je nach Technologie ergeben sich mehrere Millionen Sequenzen, sodass der direkte Vergleich mit Referenzdatenbanken unpraktikabel wird. Man verwendet daher oft einen Zwischenschritt: Über ein Clustering-Verfahren werden innerhalb eines Datensatzes Sequenzen nach Ähnlichkeit in taxonomische Einheiten (OTUs, *operational taxonomic units*) zusammengefasst (**Abb. 1D**). Aus diesen Einheiten wird jeweils nur eine repräsentative Sequenz mit der Referenzdatenbank verglichen. Da besonders im mikrobiellen Bereich der Anteil an unbekannten Organismen sehr groß werden kann, werden zudem Algorithmen eingesetzt, die bei fehlenden Referenzsequenzen die unbekannte Sequenz so gut wie möglich in übergeordnete Gruppen klassifizieren (z. B. Gattung, Familie, Ordnung).

Auch das Metabarcoding etablierte sich zuerst in der bakteriellen Ökologie. Komplette Gemeinschaften werden hier auf einmal erfasst, ohne die einzelnen Organismen vorher zu trennen [6]. Es bedarf auch keiner vorherigen Kultivierung der einzelnen Bakterien, welche für einen Großteil nicht praktikabel ist. Obwohl diese Methode noch sehr jung ist, hat sie schon enorm zu einem neuen Verständnis von mikrobieller Diversität und der Strukturierung von Gemeinschaften beigetragen [6]. Die Etablierung des Metabarcodings befindet sich derzeit auch für Eukaryoten im Aufwind und verspricht hier ebenso eine gute Erfassung der Biodiversität. Artgemeinschaften von Pilzen [7], Pflanzen [8] und Tieren [9] konnten über die Hochdurchsatzsequenzierung bereits erfolgreich erfasst werden und ermöglichen einen neuen Blick auf die Mechanismen der Etablierung und Strukturierung von Artgemeinschaften und Ökosystemen.

Jedoch ergeben sich durch das Metabarcoding auch neue Herausforderungen. Die Abundanzwerte stellen nicht unbedingt die tatsächliche Abundanz einer erfassten Art dar. Da die zugrunde liegende Polymerasekettenreaktion (PCR) kein linearer Prozess ist, kann es zu einer Überschätzung oder Unterschätzung kommen [5]. Hinzu kommt, dass die Biomasse zwischen den Arten variieren kann und dass diese auch unterschiedlich gut labortechnisch aufgeschlossen werden können. Beide Faktoren beeinträchtigen die Vergleichbarkeit von Abundanzen zwischen den Arten. Durch qualitativ schlechte Sequenziererergebnisse können Sequenzen fehlklassifiziert werden und damit zu einer artifiziellen Überschätzung der tatsächlichen Biodiversität führen. Von entscheidender Bedeutung für jede taxonomische Klassifizierung eines Metabarcoding-Datensatzes ist die Quantität und Qualität der zugrunde liegenden Referenzdatenbank, in welcher sich auch fehlerhafte Sequenzen befinden können, besonders bei nicht-kurierten Datenbanken [10]. Dem Großteil dieser neuen Schwierigkeiten kann durch eine akkurate bioinformatische Auswertung und diverse Korrekturmechanismen nach der Sequenzierung entgegengewirkt werden. Auch hier verspricht die Umstellung von einzelnen auf mehrere Marker Vorteile, ist derzeit jedoch analytisch schwerer umsetzbar als bei Einzelorganismen [11].

Anwendungsbereiche von DNA-Metabarcoding

Biodiversitätserfassung und Charakterisierung von Artgemeinschaften stellen einen essenziellen Bestandteil der ökologischen Forschung und des Naturschutzes dar. Die Mög-



lichkeit, Proben im Hochdurchsatz und ohne Auftrennung in einzelne Individuen prozessieren zu können, erlaubt generell eine Erhöhung der Stichproben (und damit der statistischen Sicherheit) sowie der Anzahl an durchführbaren Experimenten [11]. Es können auch für taxonomisch schlecht erfasste Gebiete und Artgruppen Studien durchgeführt werden. Zudem kann die Eingliederung in ökologische Nischen sehr feinskalig untersucht werden, da wenig Ausgangsmaterial notwendig ist (**Abb. 3**, [6]). Es ergeben sich neue Möglichkeiten des Naturschutzes durch das Metabarcoding von Umgebungs-DNA. Im aquatischen Bereich kann der Nachweis bedrohter oder invasiver Arten durch abgestoßene Hautzellen, Exkremente oder andere Körperbestandteile direkt über das Wasser erfolgen, ohne dass Individuen gefangen werden müssen. Durch die Beprobung von Erdschichten können Rückschlüsse auf die Biodiversität im Verlauf der Erdgeschichte gezogen werden [11].

Metabarcoding wird zudem sehr erfolgreich bei der Erfassung von zwischenartlichen Interaktionen sowie zur Identifikation von Pathogenen und Symbionten eingesetzt [6]. Es können Netzwerke aus Pflanzen und deren Bestäubern direkt über die Sequenzierung von Pollen erfasst werden [8]. Die Bedeutung von bakteriellen Gemeinschaften im Darmtrakt für die Immunabwehr und die Nährstoffversorgung ist bekannt, doch bietet die neue Forschungsmethode nun die Möglichkeit, diese Gemeinschaften systematisch zu untersuchen und im Kontext diverser Hintergründe (z. B. Ernährung und Krankheiten) auszuwerten.

◀ **Abb. 3:** Feinskalige Analyse von Bakteriengemeinschaften auf Blüten. Die Datenpunkte entsprechen Einzelproben mit jeweils einer gesamten Artgemeinschaft, aufgetragen nach Ihrer Ähnlichkeit zueinander mittels DCA(*detrended correspondence analysis*)-Ordinationsanalyse. Mikrohabitate wie Griffel, Nektar- und Staubblätter einer Blüte sowie die Blätter stellen sehr unterschiedliche Voraussetzungen für Bakterien dar. Sie beherbergen dadurch mehrere verschiedene, diverse und gut unterscheidbare Gemeinschaften, die ohne Metabarcoding bisher unterschätzt wurden (nach [6]).

Für die Sicherung des Lebensqualität der Menschen kann das Metabarcoding in einer Vielzahl von Bereichen eingesetzt werden [11]. Die Nahrungsqualität kann durch die Erfassung der pflanzlichen und tierischen Bestandteile überprüft und gesichert werden. Allergene wie Pollen in der Luft sowie Blüteereignisse bei Algen können frühzeitig erfasst und damit präventive Maßnahmen eingeleitet werden. Das Metabarcoding kann außerdem zur Erfassung von Krankheitserregern verwendet werden und damit zur Hygiene in Städten und Verkehrszentren beitragen. Kliniken sowie wissenschaftliche Labore können durch regelmäßige Prüfung auf Kontaminationen hin untersucht werden. Auch forensische Analysen lassen sich durch die Methode verbessern, indem Algen, Pollen und weitere Pflanzenbestandteile zur Ursprungsermittlung herangezogen werden.

Die Bandbreite an Applikationen ist groß, und durch die anhaltenden technologischen Weiterentwicklungen wird sowohl die Qualität als auch die Quantität der Daten durch Metabarcoding ständig verbessert und kosteneffizienter gestaltet. Mit dieser Entwicklung zeigt sich auch ein Trend in der Ausbildung der Wissenschaftler, von taxonomischen Experten hin zu bioinformatischen Analytikern. Diese verschiedenen Blickwinkel, von Metabarcoding und traditionellen Erfassungsmethoden zusammen, erlauben es, unser Wissen über Biodiversität und Artgemeinschaften deutlich zu erweitern und die Mechanismen hinter Ökosystemen zu verstehen. ■

Literatur

[1] Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

[2] Fox GE,Pechman KR, Woese CR (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int J Syst Evol Microbiol 27:44–57

[3] Hebert PD, Cywinska A, Ball SL et al. (2003) Biological identifications through DNA barcodes. Proc Biol Sci 270:313–321

[4] Dupuis J, Row A, Sperling F (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Mol Ecol 21:4422–4436

[5] Keller A, Danner N, Grimmer G et al. (2015) Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. Plant Biol 17:558–566

[6] Junker RR, Keller A (2015) Microhabitat heterogeneity across leaves and flower organs promotes bacterial diversity. FEMS Microbiol Ecol 91:fiv097 (doi: 10.1093/femsec/fiv097)

[7] Bálint M, Schmidt P, Sharma R et al. (2014) An Illumina metabarcoding pipeline for fungi. Ecol Evol 4:2642–2653

[8] Sickel W, Ankenbrand M, Grimmer G et al. (2015) Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. BMC Ecol 15:20

[9] Yu D, Ji Y, Emerson B et al. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623

[10] Nilsson R, Ryberg M, Kristiansson E et al. (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS One 1:e59

[11] Bell K, de Vere N, Keller A et al. (2016) Pollen DNA barcoding: current applications and future prospects. Genome (im Druck)

Korrespondenzadresse:
Dr. Alexander Keller
Lehrstuhl für Tierökologie und Tropenbiologie,
Biozentrum
Universität Würzburg
Am Hubland
D-97074 Würzburg
Tel.: 0931-31-84427
Fax: 0931-31-84352
a.keller@biozentrum.uni-wuerzburg.de

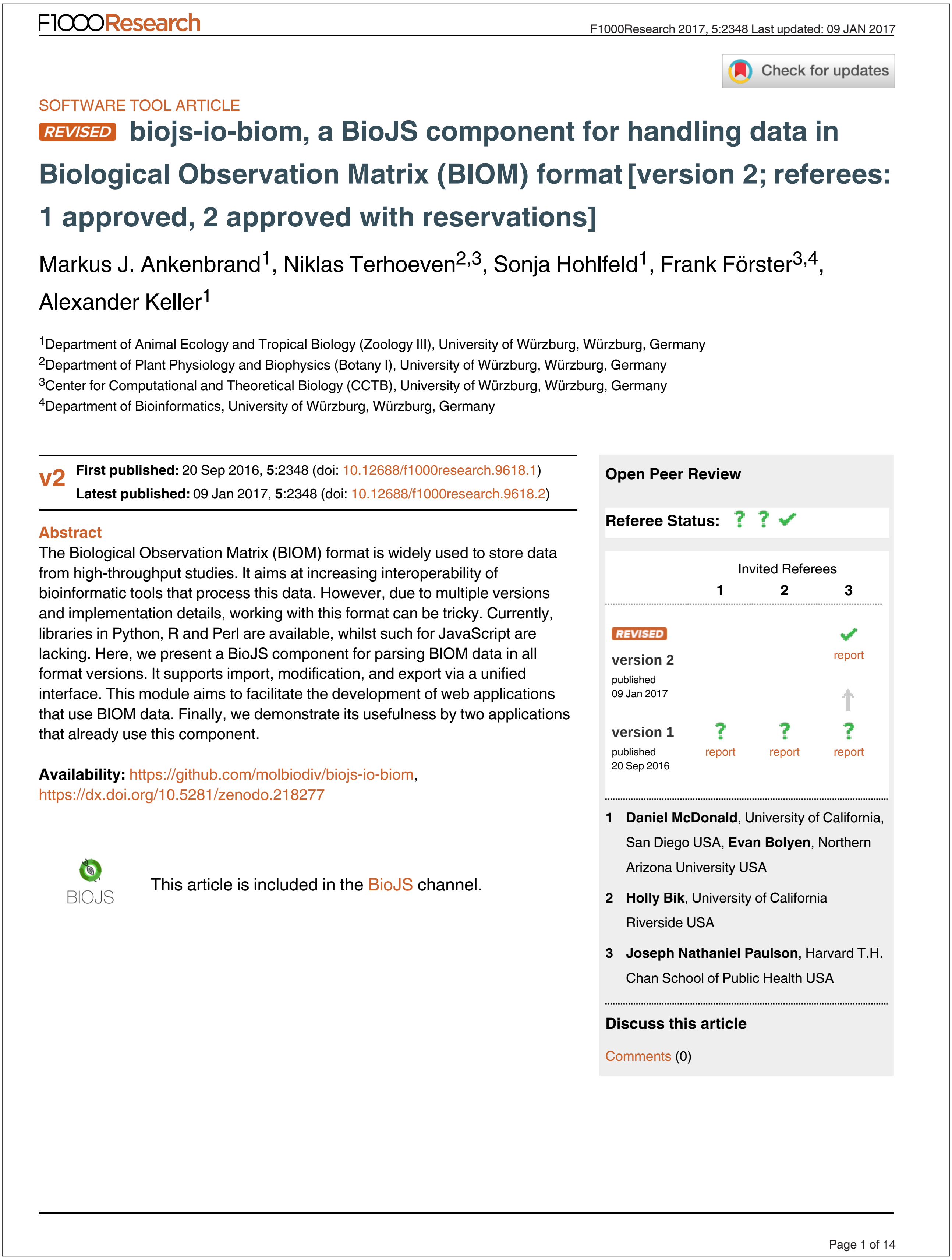
ARBEITSGRUPPE

Alexander Keller, Gudrun Grimmer, Wiebke Sickel und Markus J. Ankenbrand (v. l. n. r.)

Die *Molecular Biodiversity Group* der Universität Würzburg ist eine Nachwuchsforscher-AG, die sich mit ökologischen Artgemeinschaften beschäftigt. Wir analysieren Wirt-Mikroben-Interaktionen von Pflanzen und Insekten sowie Pflanzen-Bestäuber-Interaktionen. Zudem sind wir in der Methodenentwicklung für Metabarcoding aktiv.

5.6 BIOJS-IO-BIOM, A BIOJS COMPONENT FOR HANDLING DATA IN BIOLOGICAL OBSERVATION MATRIX (BIOM) FORMAT

– published in *F1000 Research* –



F1000Research

F1000Research 2017, 5:2348 Last updated: 09 JAN 2017

Check for updates

SOFTWARE TOOL ARTICLE

REVISED biojs-io-biom, a BioJS component for handling data in Biological Observation Matrix (BIOM) format [version 2; referees: 1 approved, 2 approved with reservations]

Markus J. Ankenbrand[1], Niklas Terhoeven[2,3], Sonja Hohlfeld[1], Frank Förster[3,4], Alexander Keller[1]

[1]Department of Animal Ecology and Tropical Biology (Zoology III), University of Würzburg, Würzburg, Germany
[2]Department of Plant Physiology and Biophysics (Botany I), University of Würzburg, Würzburg, Germany
[3]Center for Computational and Theoretical Biology (CCTB), University of Würzburg, Würzburg, Germany
[4]Department of Bioinformatics, University of Würzburg, Würzburg, Germany

v2 **First published:** 20 Sep 2016, **5**:2348 (doi: 10.12688/f1000research.9618.1)
Latest published: 09 Jan 2017, **5**:2348 (doi: 10.12688/f1000research.9618.2)

Abstract

The Biological Observation Matrix (BIOM) format is widely used to store data from high-throughput studies. It aims at increasing interoperability of bioinformatic tools that process this data. However, due to multiple versions and implementation details, working with this format can be tricky. Currently, libraries in Python, R and Perl are available, whilst such for JavaScript are lacking. Here, we present a BioJS component for parsing BIOM data in all format versions. It supports import, modification, and export via a unified interface. This module aims to facilitate the development of web applications that use BIOM data. Finally, we demonstrate its usefulness by two applications that already use this component.

Availability: https://github.com/molbiodiv/biojs-io-biom, https://dx.doi.org/10.5281/zenodo.218277

BIOJS This article is included in the BioJS channel.

Open Peer Review

Referee Status: ? ? ✓

	Invited Referees 1	2	3
REVISED version 2 published 09 Jan 2017			✓ report
version 1 published 20 Sep 2016	? report	? report	? report

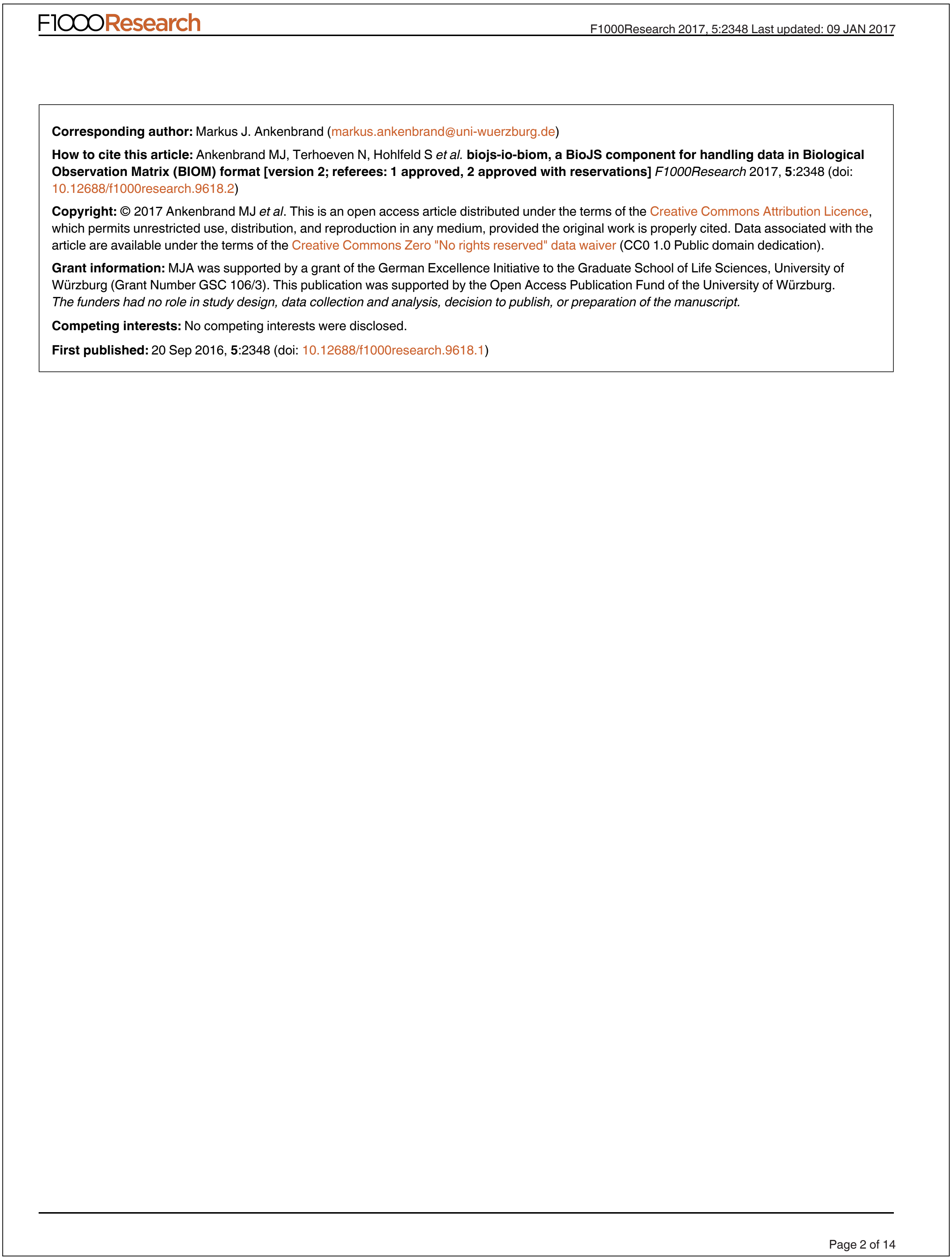
1 **Daniel McDonald**, University of California, San Diego USA, **Evan Bolyen**, Northern Arizona University USA

2 **Holly Bik**, University of California Riverside USA

3 **Joseph Nathaniel Paulson**, Harvard T.H. Chan School of Public Health USA

Discuss this article

Comments (0)



Corresponding author: Markus J. Ankenbrand (markus.ankenbrand@uni-wuerzburg.de)

How to cite this article: Ankenbrand MJ, Terhoeven N, Hohlfeld S *et al.* **biojs-io-biom, a BioJS component for handling data in Biological Observation Matrix (BIOM) format [version 2; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2017, **5**:2348 (doi: 10.12688/f1000research.9618.2)

Grant information: MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg (Grant Number GSC 106/3). This publication was supported by the Open Access Publication Fund of the University of Würzburg.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 20 Sep 2016, **5**:2348 (doi: 10.12688/f1000research.9618.1)

REVISED **Amendments from Version 1**

We added the historical context to the introduction. Further the drawbacks of relying on JSON as well as the complications with HDF5 are discussed in more detail. The application of our module to enhance Phinch now refers to a pull request into the original project rather than a fork of that. Thanks to referees comments we were able to make many small improvements (e.g. phrasing, version numbers, references).

See referee reports

Introduction

In recent years, there has been an enormous increase in biological data available from high-throughput studies. Complications arise from the enlarged size of the resulting data tables. This is the case for transcriptomic and marker-gene community data, where the central matrix consists of counts for each observation (e.g. gene or taxon) in each sample, plus a second and third matrix for metadata of both taxa and samples, respectively.

Early on there have been efforts to define data formats that capture all relevant information for an experiment like the Minimum Information About a Microarray Experiment (MIAME) project[1]. In 2005 the Genomic Standards Consortium (GSC) formed with the mission of enabling genomic data integration, discovery and comparison through international community-driven standards[2]. The Biological Observation Matrix (BIOM) Format was developed to standardize the storage of observation counts together with all relevant metadata and it is a member project of the GSC[3]. One main purpose of the BIOM format is to enhance interoperability between different software suits. Many current leading tools in community ecology and metagenomics support the BIOM format, e.g. QIIME[4], MG-RAST[5], PICRUSt[6], phyloseq[7], VAMPS[8] and Phinch[9]. Additionally, libraries exist in Python[3], R[10] and Perl[11] to propagate the standardized use of the format.

Interactive visualization of biological data in a web browser is becoming more and more popular[12,13]. For the development of web applications that support BIOM data, a corresponding library is currently lacking and would be very useful, since several challenges arise when trying to handle BIOM data. While BIOM format version 1.0 builds on the JSON format and thus is natively supported by JavaScript, the more recent BIOM format version 2.1 uses HDF5 and can therefore not be handled natively in web browsers. Also the internal data storage can be either dense or sparse so applications have to handle both cases. Furthermore application developers need to be very careful when modifying BIOM data as changes that do not abide to the specification will break interoperability with other tools. Here we present biojs-io-biom, a JavaScript module that provides a unified interface to read, modify, and write BIOM data. It can be readily used as a library by applications that need to handle BIOM data for import or export directly in the browser. To demonstrate the utility of our module it has been used to implement a simple user interface for the biom-conversion-server[14]. Additionally, the popular BIOM visualization tool Phinch[9] has been extended with new features, in particular support for BIOM version 2.1 by integrating biojs-io-biom[15].

The biojs-io-biom component

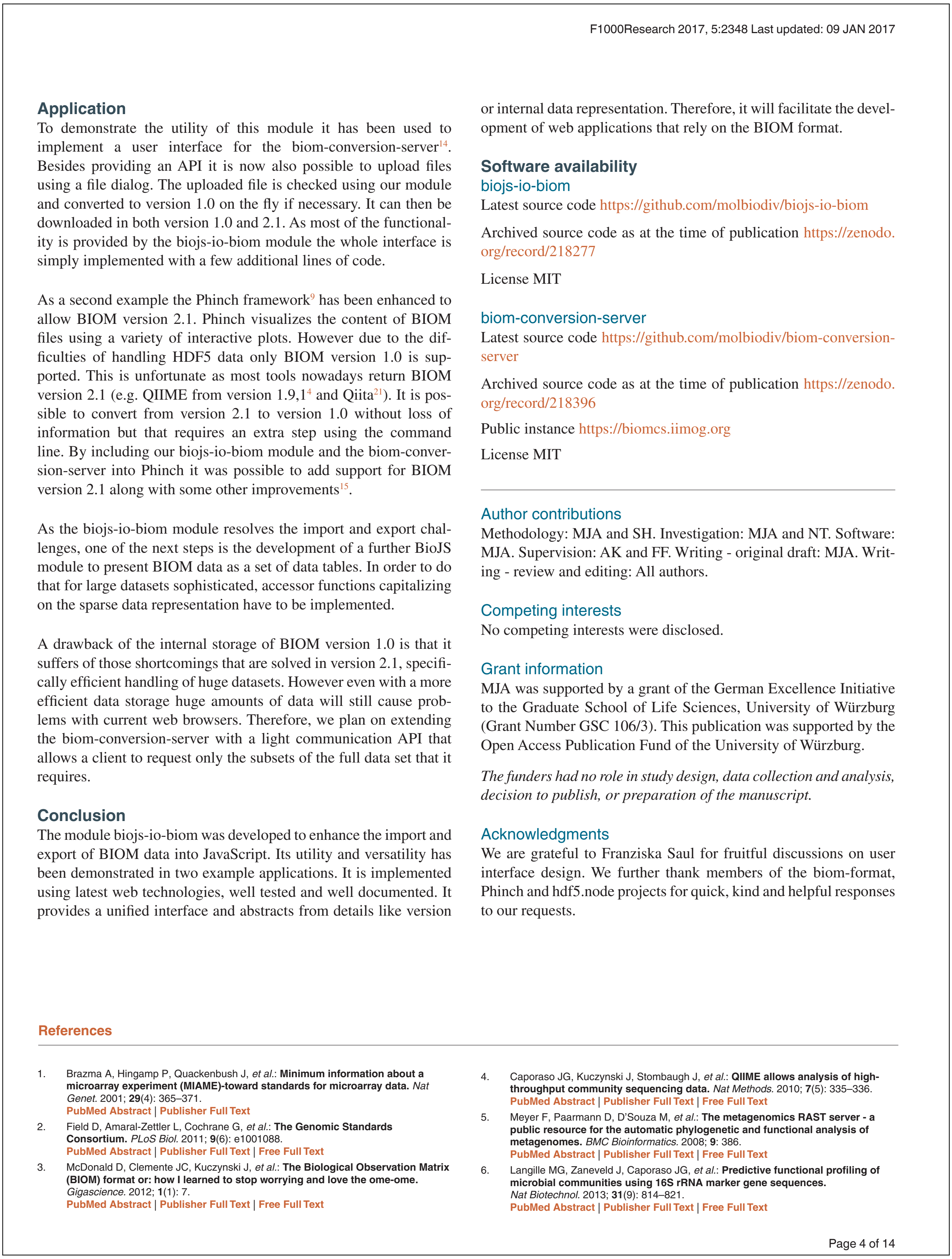
The biojs-io-biom library can be used to create new objects (called `Biom` objects for brevity) by either loading file content directly via the static `parse` function or by initialization with a JSON object:

```
var biom = new Biom({
    id: 'My Biom',
    matrix_type: 'dense',
    shape: [2,2],
    rows: [
        {id: 'row1', metadata: {}},
        {id: 'row2', metadata: {}}
    ],
    columns: [
        {id: 'col1', metadata: {}},
        {id: 'col2', metadata: {}}
    ],
    data: [
        [0,1],
        [2,3]
    ]
});
```

The data is checked for integrity and compliance with the BIOM specification. Missing fields are created with default content. All operations that set attributes of the `Biom` object with the dot notation are also checked and prompt an error if they are not allowed.

```
var biom = new Biom({});
biom.id = [];
// Will throw a TypeError as id has to be a
string or null
```

Beside checking and maintaining integrity the biojs-io-biom library implements convenience functions. This includes getter and setter for metadata as well as data accessor functions that are agnostic to internal representation (dense or sparse). But one of the main features of this library is the capability of handling BIOM data in both versions 1.0 and 2.1 by interfacing with the biom-conversion-server[14]. Handling of BIOM version 2.1 in JavaScript directly is not possible due to its HDF5 binary format. The only reference implementation of the format is in C and trying to transpile the library to JavaScript using emscripten[16] failed due to strong reliance on fle operations (see discussions in[17,18]). Using the conversion server allows developers to use BIOM of both versions transparently. `Biom` objects also expose the function `write` which exports it as version 1.0 or version 2.1. In contrast to the existing `biom_convert` module for the Galaxy platform which has a rich set of options the biom-conversion-server exhibits its functionality both via an API and a simple user interface that does not need any kind of setup or login[19,20].



Application

To demonstrate the utility of this module it has been used to implement a user interface for the biom-conversion-server[14]. Besides providing an API it is now also possible to upload files using a file dialog. The uploaded file is checked using our module and converted to version 1.0 on the fly if necessary. It can then be downloaded in both version 1.0 and 2.1. As most of the functionality is provided by the biojs-io-biom module the whole interface is simply implemented with a few additional lines of code.

As a second example the Phinch framework[9] has been enhanced to allow BIOM version 2.1. Phinch visualizes the content of BIOM files using a variety of interactive plots. However due to the difficulties of handling HDF5 data only BIOM version 1.0 is supported. This is unfortunate as most tools nowadays return BIOM version 2.1 (e.g. QIIME from version 1.9,1[4] and Qiita[21]). It is possible to convert from version 2.1 to version 1.0 without loss of information but that requires an extra step using the command line. By including our biojs-io-biom module and the biom-conversion-server into Phinch it was possible to add support for BIOM version 2.1 along with some other improvements[15].

As the biojs-io-biom module resolves the import and export challenges, one of the next steps is the development of a further BioJS module to present BIOM data as a set of data tables. In order to do that for large datasets sophisticated, accessor functions capitalizing on the sparse data representation have to be implemented.

A drawback of the internal storage of BIOM version 1.0 is that it suffers of those shortcomings that are solved in version 2.1, specifically efficient handling of huge datasets. However even with a more efficient data storage huge amounts of data will still cause problems with current web browsers. Therefore, we plan on extending the biom-conversion-server with a light communication API that allows a client to request only the subsets of the full data set that it requires.

Conclusion

The module biojs-io-biom was developed to enhance the import and export of BIOM data into JavaScript. Its utility and versatility has been demonstrated in two example applications. It is implemented using latest web technologies, well tested and well documented. It provides a unified interface and abstracts from details like version or internal data representation. Therefore, it will facilitate the development of web applications that rely on the BIOM format.

Software availability

biojs-io-biom

Latest source code https://github.com/molbiodiv/biojs-io-biom

Archived source code as at the time of publication https://zenodo.org/record/218277

License MIT

biom-conversion-server

Latest source code https://github.com/molbiodiv/biom-conversion-server

Archived source code as at the time of publication https://zenodo.org/record/218396

Public instance https://biomcs.iimog.org

License MIT

Author contributions

Methodology: MJA and SH. Investigation: MJA and NT. Software: MJA. Supervision: AK and FF. Writing - original draft: MJA. Writing - review and editing: All authors.

Competing interests

No competing interests were disclosed.

Grant information

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg (Grant Number GSC 106/3). This publication was supported by the Open Access Publication Fund of the University of Würzburg.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We are grateful to Franziska Saul for fruitful discussions on user interface design. We further thank members of the biom-format, Phinch and hdf5.node projects for quick, kind and helpful responses to our requests.

References

1. Brazma A, Hingamp P, Quackenbush J, *et al.*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet.* 2001; **29**(4): 365–371.
PubMed Abstract | Publisher Full Text
2. Field D, Amaral-Zettler L, Cochrane G, *et al.*: **The Genomic Standards Consortium.** *PLoS Biol.* 2011; **9**(6): e1001088.
PubMed Abstract | Publisher Full Text | Free Full Text
3. McDonald D, Clemente JC, Kuczynski J, *et al.*: **The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome.** *Gigascience.* 2012; **1**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Caporaso JG, Kuczynski J, Stombaugh J, *et al.*: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods.* 2010; **7**(5): 335–336.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Meyer F, Paarmann D, D'Souza M, *et al.*: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.** *BMC Bioinformatics.* 2008; **9**: 386.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Langille MG, Zaneveld J, Caporaso JG, *et al.*: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.** *Nat Biotechnol.* 2013; **31**(9): 814–821.
PubMed Abstract | Publisher Full Text | Free Full Text

7. McMurdie PJ, Holmes S: **Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PLoS One.* 2013; **8**(4): e61217.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Huse SM, Mark Welch DB, Voorhis A, *et al.*: **VAMPS: a website for visualization and analysis of microbial population structures.** *BMC Bioinformatics.* 2014; **15**: 41.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Bik HM; Pitch Interactive: **Phinch: An interactive, exploratory data visualization framework for –Omic datasets.** *bioRxiv.* 2014; 009944.
Publisher Full Text

10. McMurdie PJ, Paulson JN: **biomformat: An interface package for the BIOM file format.** R/Bioconductor package version 1.0.0. 2015.

11. Angly FE, Fields CJ, Tyson GW: **The Bio-Community Perl toolkit for microbial ecology.** *Bioinformatics.* 2014; **30**(13): 1926–1927.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Corpas M, Jimenez R, Carbon SJ, *et al.*: **BioJS: an open source standard for biological visualisation - its status in 2014 [version 1; referees: 2 approved].** *F1000Res.* 2014; **3**: 55.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Corpas M: **The BioJS article collection of open source components for biological data visualisation [version 1; referees: not peer reviewed].** *F1000Res.* 2014; **3**: 56.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Ankenbrand MJ: **molbiodiv/biom-conversion-server: Version 1.0.2**. 2016.
Publisher Full Text

15. **Pull request #67· PitchInteractiveInc/Phinch.** preview version online at https://blackbird.iimog.org. Accessed: 2016-12-22.
Reference Source

16. Kripken/emscripten: **Emscripten: An LLVM-to-JavaScript Compiler.** Accessed: 2016-09-08.
Reference Source

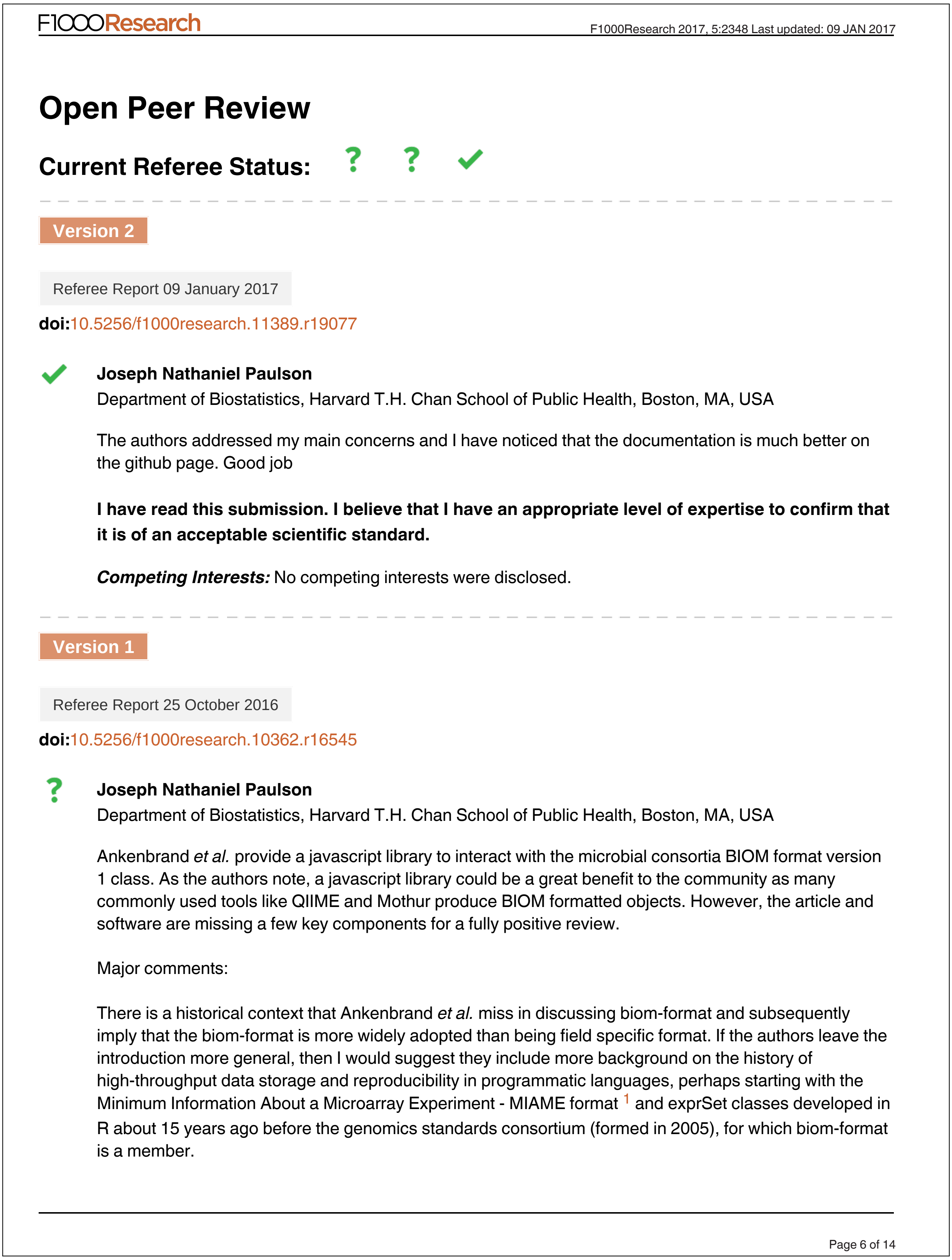
17. **Biom javascript module· Issue #699· biocore/biom-format.** Accessed: 2016-09-08.
Reference Source

18. **hdf5 javascript in a webbrowser· Issue #29· HDF-NI/hdf5.node.** Accessed: 2016-09-08.
Reference Source

19. Afgan E, Baker D, van den Beek M, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.** *Nucleic Acids Res.* 2016; **44**(W1): W3–W10.
PubMed Abstract | Publisher Full Text | Free Full Text

20. **biom convert galaxy module.** Accessed: 2016-12-15.
Reference Source

21. **Qiita.** Accessed: 2016-09-08.
Reference Source



Open Peer Review

Current Referee Status: ? ? ✓

Version 2

Referee Report 09 January 2017

doi:10.5256/f1000research.11389.r19077

✓ **Joseph Nathaniel Paulson**
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

The authors addressed my main concerns and I have noticed that the documentation is much better on the github page. Good job

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 25 October 2016

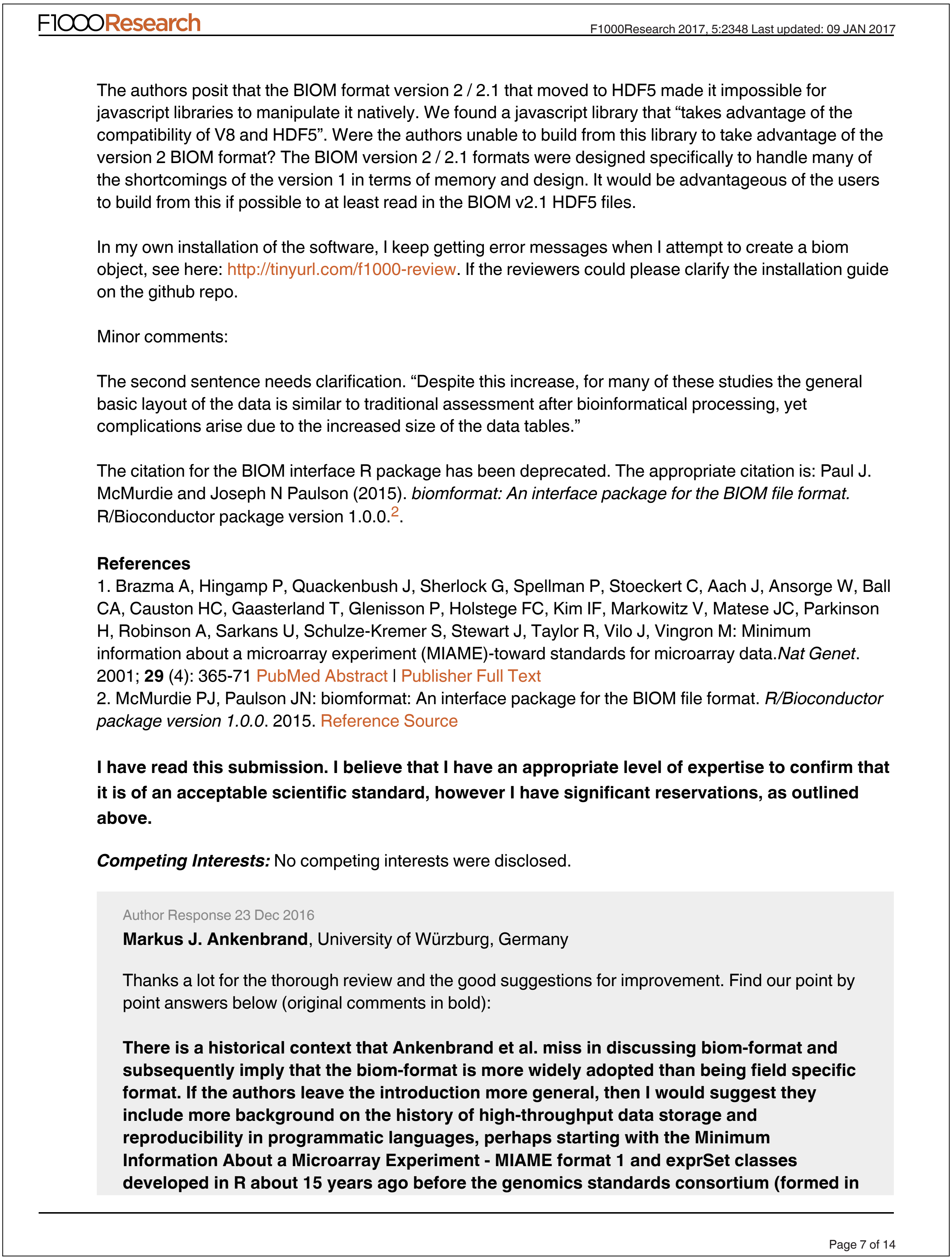
doi:10.5256/f1000research.10362.r16545

? **Joseph Nathaniel Paulson**
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Ankenbrand *et al.* provide a javascript library to interact with the microbial consortia BIOM format version 1 class. As the authors note, a javascript library could be a great benefit to the community as many commonly used tools like QIIME and Mothur produce BIOM formatted objects. However, the article and software are missing a few key components for a fully positive review.

Major comments:

There is a historical context that Ankenbrand *et al.* miss in discussing biom-format and subsequently imply that the biom-format is more widely adopted than being field specific format. If the authors leave the introduction more general, then I would suggest they include more background on the history of high-throughput data storage and reproducibility in programmatic languages, perhaps starting with the Minimum Information About a Microarray Experiment - MIAME format [1] and exprSet classes developed in R about 15 years ago before the genomics standards consortium (formed in 2005), for which biom-format is a member.



F1000Research F1000Research 2017, 5:2348 Last updated: 09 JAN 2017

The authors posit that the BIOM format version 2 / 2.1 that moved to HDF5 made it impossible for javascript libraries to manipulate it natively. We found a javascript library that “takes advantage of the compatibility of V8 and HDF5”. Were the authors unable to build from this library to take advantage of the version 2 BIOM format? The BIOM version 2 / 2.1 formats were designed specifically to handle many of the shortcomings of the version 1 in terms of memory and design. It would be advantageous of the users to build from this if possible to at least read in the BIOM v2.1 HDF5 files.

In my own installation of the software, I keep getting error messages when I attempt to create a biom object, see here: http://tinyurl.com/f1000-review. If the reviewers could please clarify the installation guide on the github repo.

Minor comments:

The second sentence needs clarification. “Despite this increase, for many of these studies the general basic layout of the data is similar to traditional assessment after bioinformatical processing, yet complications arise due to the increased size of the data tables.”

The citation for the BIOM interface R package has been deprecated. The appropriate citation is: Paul J. McMurdie and Joseph N Paulson (2015). *biomformat: An interface package for the BIOM file format.* R/Bioconductor package version 1.0.0.[2].

References
1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.*Nat Genet*. 2001; **29** (4): 365-71 PubMed Abstract | Publisher Full Text
2. McMurdie PJ, Paulson JN: biomformat: An interface package for the BIOM file format. *R/Bioconductor package version 1.0.0*. 2015. Reference Source

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

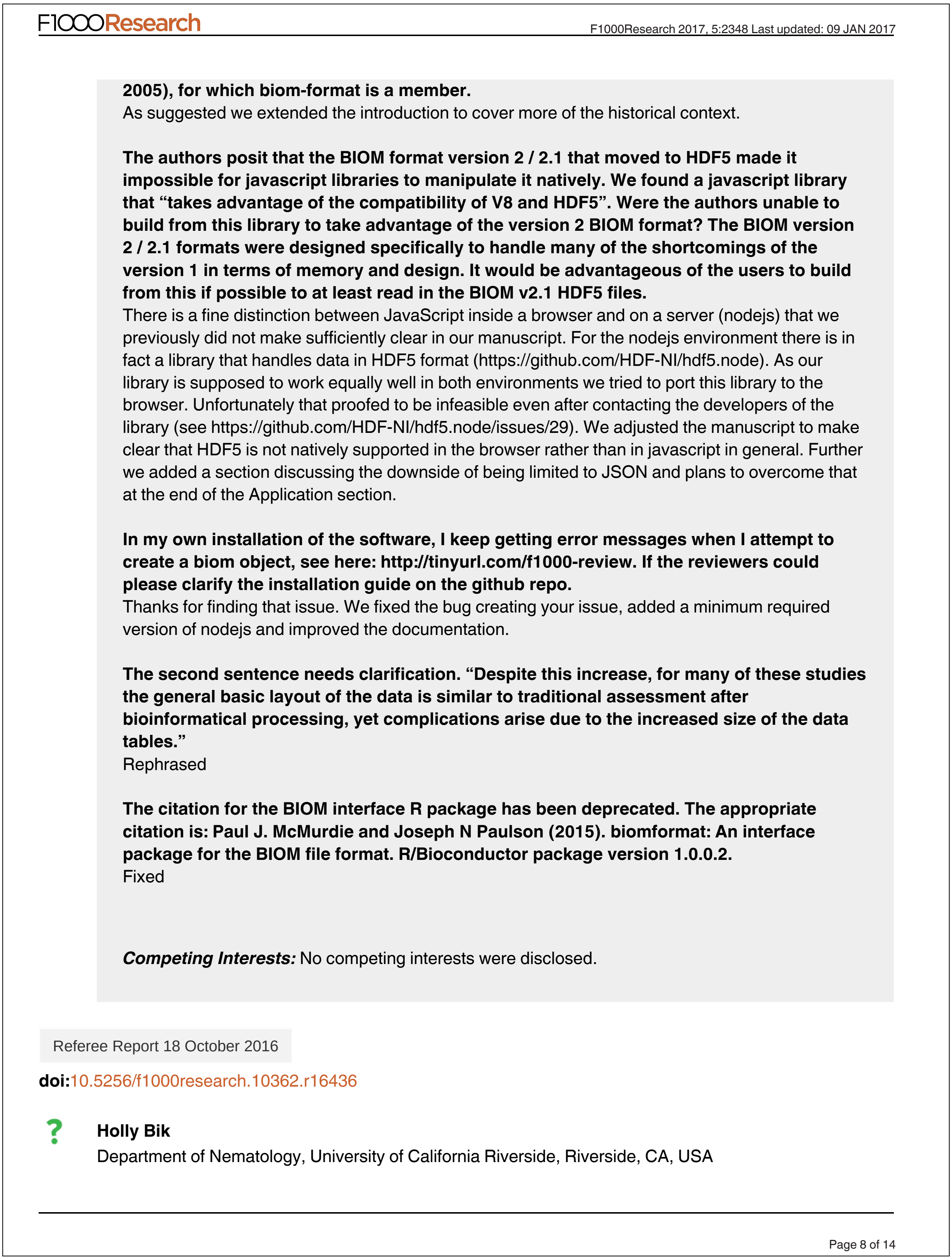
Author Response 23 Dec 2016

Markus J. Ankenbrand, University of Würzburg, Germany

Thanks a lot for the thorough review and the good suggestions for improvement. Find our point by point answers below (original comments in bold):

There is a historical context that Ankenbrand et al. miss in discussing biom-format and subsequently imply that the biom-format is more widely adopted than being field specific format. If the authors leave the introduction more general, then I would suggest they include more background on the history of high-throughput data storage and reproducibility in programmatic languages, perhaps starting with the Minimum Information About a Microarray Experiment - MIAME format 1 and exprSet classes developed in R about 15 years ago before the genomics standards consortium (formed in

Page 7 of 14



2005), for which biom-format is a member.
As suggested we extended the introduction to cover more of the historical context.

The authors posit that the BIOM format version 2 / 2.1 that moved to HDF5 made it impossible for javascript libraries to manipulate it natively. We found a javascript library that “takes advantage of the compatibility of V8 and HDF5”. Were the authors unable to build from this library to take advantage of the version 2 BIOM format? The BIOM version 2 / 2.1 formats were designed specifically to handle many of the shortcomings of the version 1 in terms of memory and design. It would be advantageous of the users to build from this if possible to at least read in the BIOM v2.1 HDF5 files.
There is a fine distinction between JavaScript inside a browser and on a server (nodejs) that we previously did not make sufficiently clear in our manuscript. For the nodejs environment there is in fact a library that handles data in HDF5 format (https://github.com/HDF-NI/hdf5.node). As our library is supposed to work equally well in both environments we tried to port this library to the browser. Unfortunately that proofed to be infeasible even after contacting the developers of the library (see https://github.com/HDF-NI/hdf5.node/issues/29). We adjusted the manuscript to make clear that HDF5 is not natively supported in the browser rather than in javascript in general. Further we added a section discussing the downside of being limited to JSON and plans to overcome that at the end of the Application section.

In my own installation of the software, I keep getting error messages when I attempt to create a biom object, see here: http://tinyurl.com/f1000-review. If the reviewers could please clarify the installation guide on the github repo.
Thanks for finding that issue. We fixed the bug creating your issue, added a minimum required version of nodejs and improved the documentation.

The second sentence needs clarification. “Despite this increase, for many of these studies the general basic layout of the data is similar to traditional assessment after bioinformatical processing, yet complications arise due to the increased size of the data tables.”
Rephrased

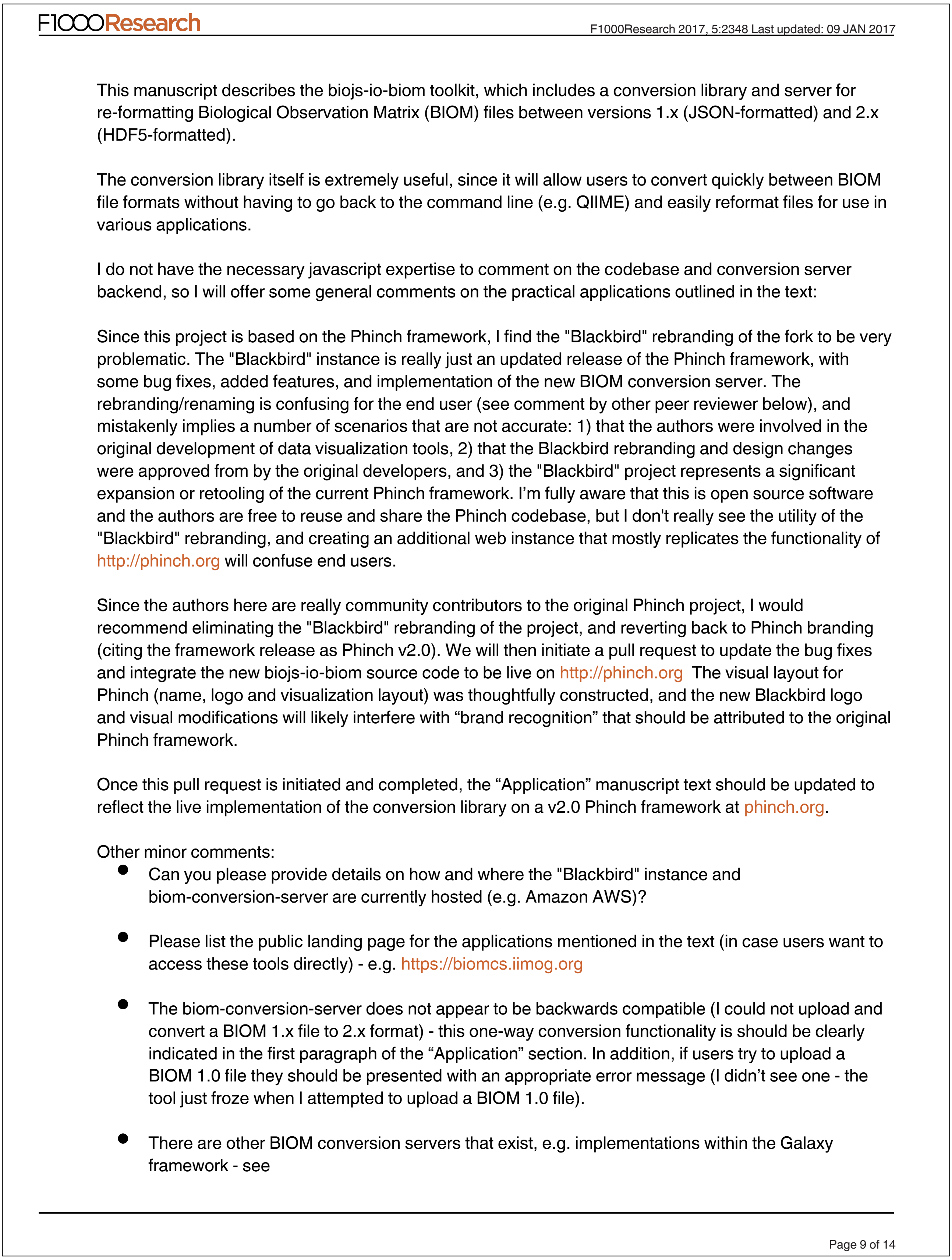
The citation for the BIOM interface R package has been deprecated. The appropriate citation is: Paul J. McMurdie and Joseph N Paulson (2015). biomformat: An interface package for the BIOM file format. R/Bioconductor package version 1.0.0.2.
Fixed

Competing Interests: No competing interests were disclosed.

Referee Report 18 October 2016

doi:10.5256/f1000research.10362.r16436

? **Holly Bik**
Department of Nematology, University of California Riverside, Riverside, CA, USA



This manuscript describes the biojs-io-biom toolkit, which includes a conversion library and server for re-formatting Biological Observation Matrix (BIOM) files between versions 1.x (JSON-formatted) and 2.x (HDF5-formatted).

The conversion library itself is extremely useful, since it will allow users to convert quickly between BIOM file formats without having to go back to the command line (e.g. QIIME) and easily reformat files for use in various applications.

I do not have the necessary javascript expertise to comment on the codebase and conversion server backend, so I will offer some general comments on the practical applications outlined in the text:

Since this project is based on the Phinch framework, I find the "Blackbird" rebranding of the fork to be very problematic. The "Blackbird" instance is really just an updated release of the Phinch framework, with some bug fixes, added features, and implementation of the new BIOM conversion server. The rebranding/renaming is confusing for the end user (see comment by other peer reviewer below), and mistakenly implies a number of scenarios that are not accurate: 1) that the authors were involved in the original development of data visualization tools, 2) that the Blackbird rebranding and design changes were approved from by the original developers, and 3) the "Blackbird" project represents a significant expansion or retooling of the current Phinch framework. I'm fully aware that this is open source software and the authors are free to reuse and share the Phinch codebase, but I don't really see the utility of the "Blackbird" rebranding, and creating an additional web instance that mostly replicates the functionality of http://phinch.org will confuse end users.

Since the authors here are really community contributors to the original Phinch project, I would recommend eliminating the "Blackbird" rebranding of the project, and reverting back to Phinch branding (citing the framework release as Phinch v2.0). We will then initiate a pull request to update the bug fixes and integrate the new biojs-io-biom source code to be live on http://phinch.org The visual layout for Phinch (name, logo and visualization layout) was thoughtfully constructed, and the new Blackbird logo and visual modifications will likely interfere with "brand recognition" that should be attributed to the original Phinch framework.

Once this pull request is initiated and completed, the "Application" manuscript text should be updated to reflect the live implementation of the conversion library on a v2.0 Phinch framework at phinch.org.

Other minor comments:

- Can you please provide details on how and where the "Blackbird" instance and biom-conversion-server are currently hosted (e.g. Amazon AWS)?
- Please list the public landing page for the applications mentioned in the text (in case users want to access these tools directly) - e.g. https://biomcs.iimog.org
- The biom-conversion-server does not appear to be backwards compatible (I could not upload and convert a BIOM 1.x file to 2.x format) - this one-way conversion functionality is should be clearly indicated in the first paragraph of the "Application" section. In addition, if users try to upload a BIOM 1.0 file they should be presented with an appropriate error message (I didn't see one - the tool just froze when I attempted to upload a BIOM 1.0 file).
- There are other BIOM conversion servers that exist, e.g. implementations within the Galaxy framework - see

https://toolshed.g2.bx.psu.edu/repository/display_tool?repository_id=b3ae8ca9317b000e&render_
- these alternate tools should be mentioned in the text. How does the biom-conversion-server compare with (and potentially improve on) such Galaxy based tools?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

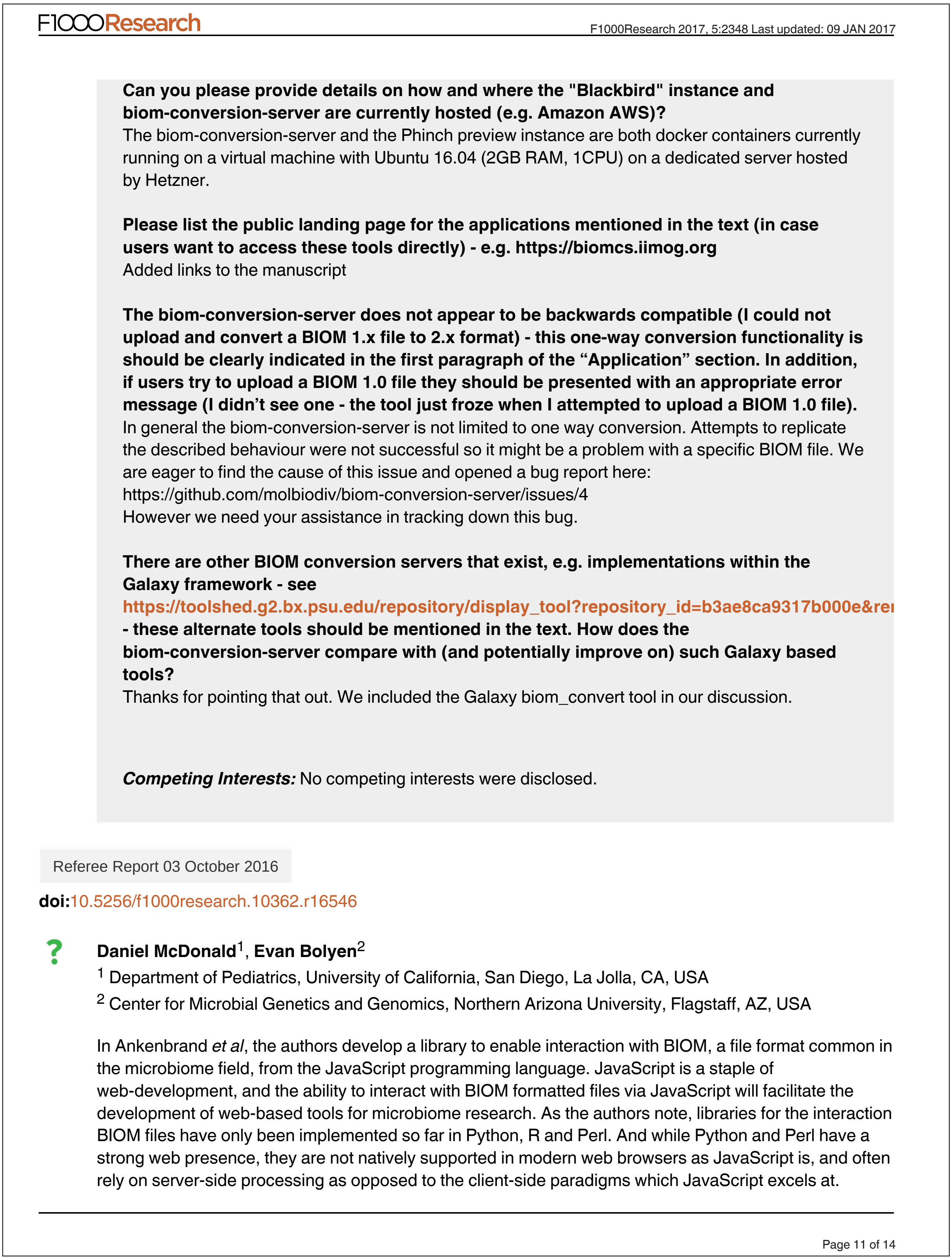
Competing Interests: I am the Principal Investigator on the Phinch framework (http://phinch.org) which is the underlying codebase used to generate the "Blackbird" application mentioned in this manuscript.

Author Response 23 Dec 2016

Markus J. Ankenbrand, University of Würzburg, Germany

Thanks a lot for taking the time to review this article and for the good suggestions for improvement. Find our point by point answers below (original comments in bold):

Since this project is based on the Phinch framework, I find the "Blackbird" rebranding of the fork to be very problematic. The "Blackbird" instance is really just an updated release of the Phinch framework, with some bug fixes, added features, and implementation of the new BIOM conversion server. The rebranding/renaming is confusing for the end user (see comment by other peer reviewer below), and mistakenly implies a number of scenarios that are not accurate:
1) that the authors were involved in the original development of data visualization tools,
2) that the Blackbird rebranding and design changes were approved from by the original developers, and
3) the "Blackbird" project represents a significant expansion or retooling of the current Phinch framework.
I'm fully aware that this is open source software and the authors are free to reuse and share the Phinch codebase, but I don't really see the utility of the "Blackbird" rebranding, and creating an additional web instance that mostly replicates the functionality of http://phinch.org will confuse end users. Since the authors here are really community contributors to the original Phinch project, I would recommend eliminating the "Blackbird" rebranding of the project, and reverting back to Phinch branding (citing the framework release as Phinch v2.0).We will then initiate a pull request to update the bug fixes and integrate the new biojs-io-biom source code to be live on http://phinch.org The visual layout for Phinch (name, logo and visualization layout) was thoughtfully constructed, and the new Blackbird logo and visual modifications will likely interfere with “brand recognition” that should be attributed to the original Phinch framework. Once this pull request is initiated and completed, the “Application” manuscript text should be updated to reflect the live implementation of the conversion library on a v2.0 Phinch framework at phinch.org.
Thanks for sharing your thoughts on this delicate topic. We are grateful to you for suggesting a more satisfactory solution. As you suggested we prepared the pull request that integrates the additional features into Phinch and removed Blackbird branding from our fork. We look forward to the changes going live on phinch.org. We will use the same procedure for future improvements as long as you are interested in merging them.



Can you please provide details on how and where the "Blackbird" instance and biom-conversion-server are currently hosted (e.g. Amazon AWS)?
The biom-conversion-server and the Phinch preview instance are both docker containers currently running on a virtual machine with Ubuntu 16.04 (2GB RAM, 1CPU) on a dedicated server hosted by Hetzner.

Please list the public landing page for the applications mentioned in the text (in case users want to access these tools directly) - e.g. https://biomcs.iimog.org
Added links to the manuscript

The biom-conversion-server does not appear to be backwards compatible (I could not upload and convert a BIOM 1.x file to 2.x format) - this one-way conversion functionality is should be clearly indicated in the first paragraph of the “Application” section. In addition, if users try to upload a BIOM 1.0 file they should be presented with an appropriate error message (I didn’t see one - the tool just froze when I attempted to upload a BIOM 1.0 file).
In general the biom-conversion-server is not limited to one way conversion. Attempts to replicate the described behaviour were not successful so it might be a problem with a specific BIOM file. We are eager to find the cause of this issue and opened a bug report here:
https://github.com/molbiodiv/biom-conversion-server/issues/4
However we need your assistance in tracking down this bug.

There are other BIOM conversion servers that exist, e.g. implementations within the Galaxy framework - see https://toolshed.g2.bx.psu.edu/repository/display_tool?repository_id=b3ae8ca9317b000e&re - these alternate tools should be mentioned in the text. How does the biom-conversion-server compare with (and potentially improve on) such Galaxy based tools?
Thanks for pointing that out. We included the Galaxy biom_convert tool in our discussion.

Competing Interests: No competing interests were disclosed.

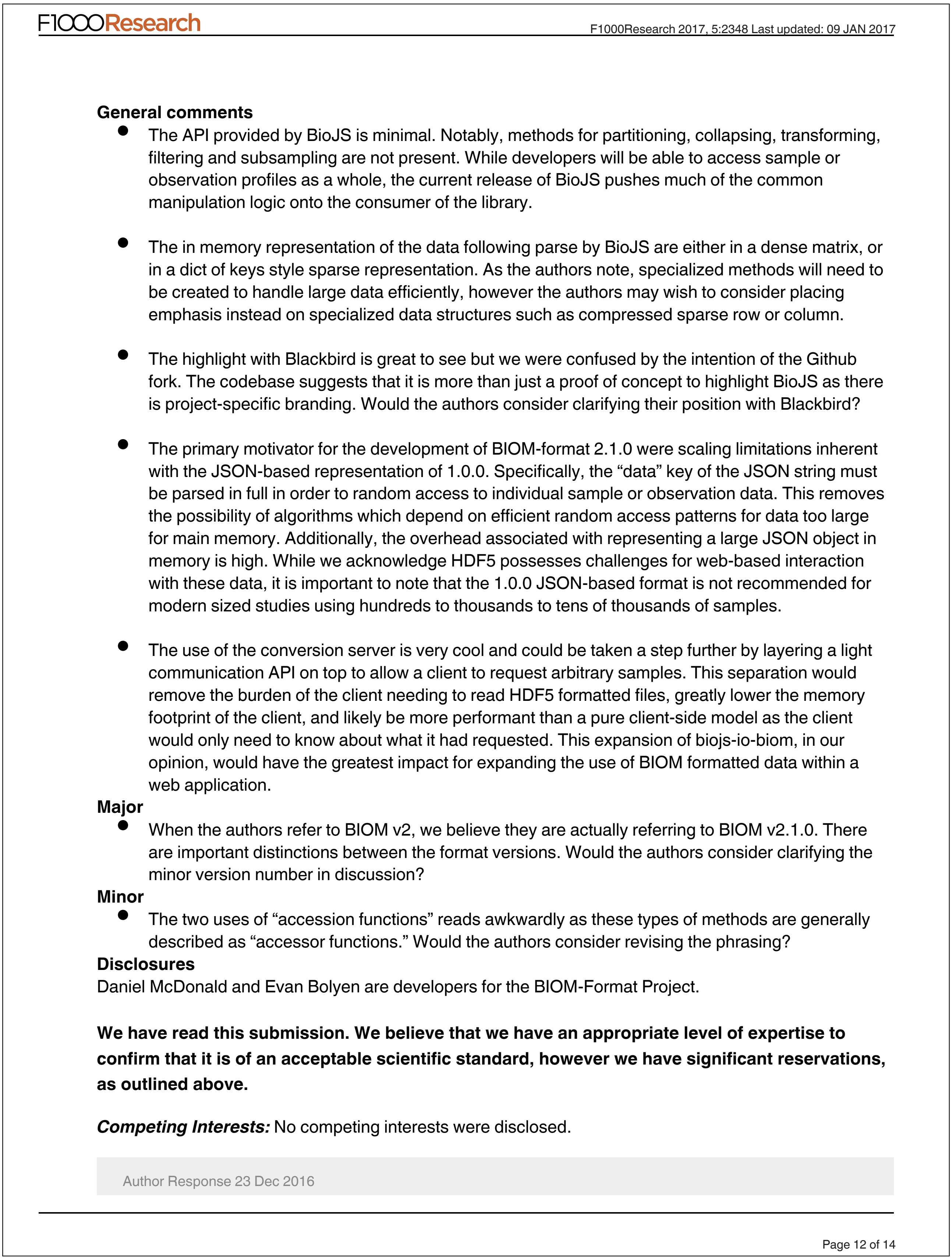
Referee Report 03 October 2016

doi:10.5256/f1000research.10362.r16546

?

Daniel McDonald[1], **Evan Bolyen**[2]
[1] Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA
[2] Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ, USA

In Ankenbrand *et al*, the authors develop a library to enable interaction with BIOM, a file format common in the microbiome field, from the JavaScript programming language. JavaScript is a staple of web-development, and the ability to interact with BIOM formatted files via JavaScript will facilitate the development of web-based tools for microbiome research. As the authors note, libraries for the interaction BIOM files have only been implemented so far in Python, R and Perl. And while Python and Perl have a strong web presence, they are not natively supported in modern web browsers as JavaScript is, and often rely on server-side processing as opposed to the client-side paradigms which JavaScript excels at.



General comments

- The API provided by BioJS is minimal. Notably, methods for partitioning, collapsing, transforming, filtering and subsampling are not present. While developers will be able to access sample or observation profiles as a whole, the current release of BioJS pushes much of the common manipulation logic onto the consumer of the library.
- The in memory representation of the data following parse by BioJS are either in a dense matrix, or in a dict of keys style sparse representation. As the authors note, specialized methods will need to be created to handle large data efficiently, however the authors may wish to consider placing emphasis instead on specialized data structures such as compressed sparse row or column.
- The highlight with Blackbird is great to see but we were confused by the intention of the Github fork. The codebase suggests that it is more than just a proof of concept to highlight BioJS as there is project-specific branding. Would the authors consider clarifying their position with Blackbird?
- The primary motivator for the development of BIOM-format 2.1.0 were scaling limitations inherent with the JSON-based representation of 1.0.0. Specifically, the “data” key of the JSON string must be parsed in full in order to random access to individual sample or observation data. This removes the possibility of algorithms which depend on efficient random access patterns for data too large for main memory. Additionally, the overhead associated with representing a large JSON object in memory is high. While we acknowledge HDF5 possesses challenges for web-based interaction with these data, it is important to note that the 1.0.0 JSON-based format is not recommended for modern sized studies using hundreds to thousands to tens of thousands of samples.
- The use of the conversion server is very cool and could be taken a step further by layering a light communication API on top to allow a client to request arbitrary samples. This separation would remove the burden of the client needing to read HDF5 formatted files, greatly lower the memory footprint of the client, and likely be more performant than a pure client-side model as the client would only need to know about what it had requested. This expansion of biojs-io-biom, in our opinion, would have the greatest impact for expanding the use of BIOM formatted data within a web application.

Major

- When the authors refer to BIOM v2, we believe they are actually referring to BIOM v2.1.0. There are important distinctions between the format versions. Would the authors consider clarifying the minor version number in discussion?

Minor

- The two uses of “accession functions” reads awkwardly as these types of methods are generally described as “accessor functions.” Would the authors consider revising the phrasing?

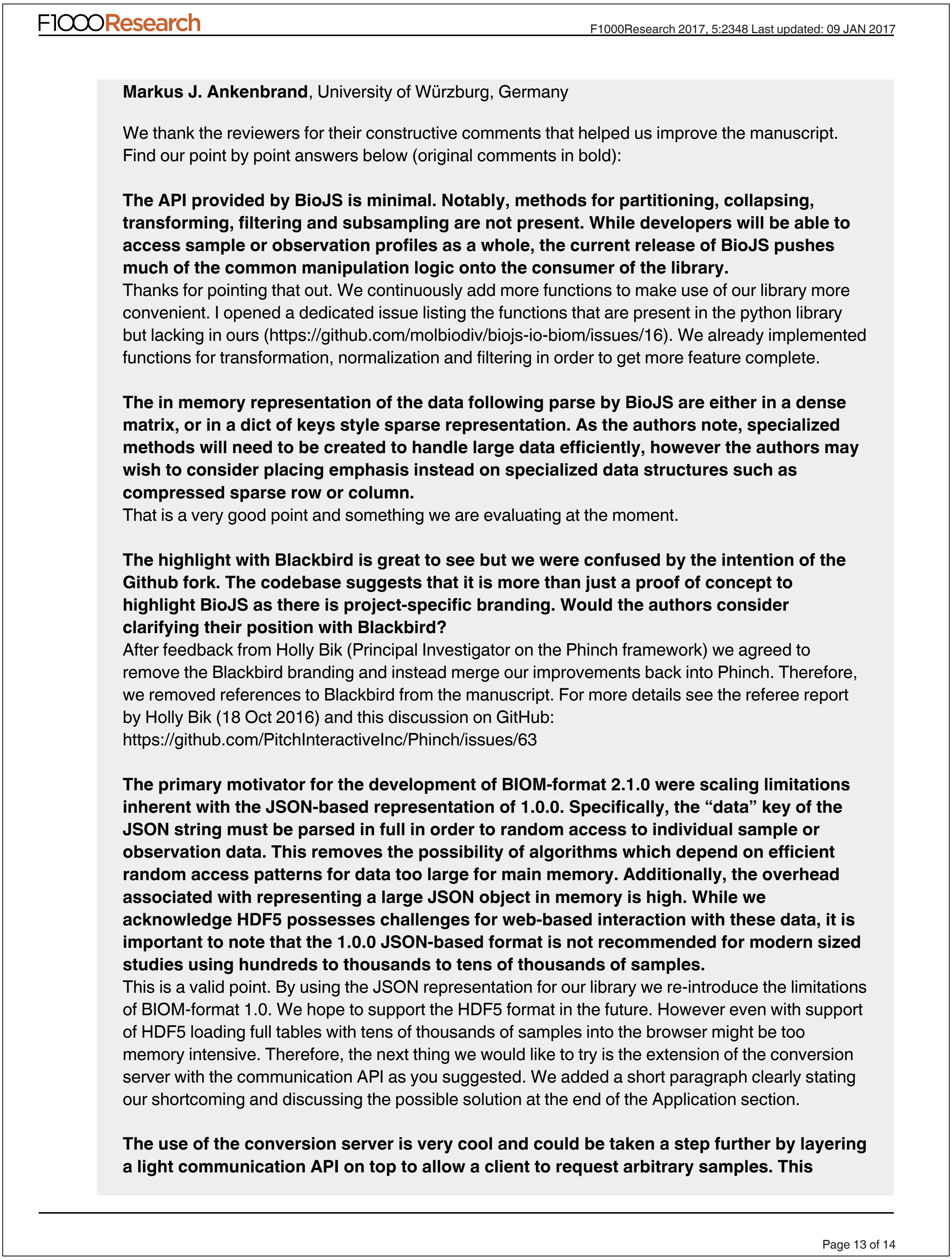
Disclosures

Daniel McDonald and Evan Bolyen are developers for the BIOM-Format Project.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 23 Dec 2016



Markus J. Ankenbrand, University of Würzburg, Germany

We thank the reviewers for their constructive comments that helped us improve the manuscript. Find our point by point answers below (original comments in bold):

The API provided by BioJS is minimal. Notably, methods for partitioning, collapsing, transforming, filtering and subsampling are not present. While developers will be able to access sample or observation profiles as a whole, the current release of BioJS pushes much of the common manipulation logic onto the consumer of the library.
Thanks for pointing that out. We continuously add more functions to make use of our library more convenient. I opened a dedicated issue listing the functions that are present in the python library but lacking in ours (https://github.com/molbiodiv/biojs-io-biom/issues/16). We already implemented functions for transformation, normalization and filtering in order to get more feature complete.

The in memory representation of the data following parse by BioJS are either in a dense matrix, or in a dict of keys style sparse representation. As the authors note, specialized methods will need to be created to handle large data efficiently, however the authors may wish to consider placing emphasis instead on specialized data structures such as compressed sparse row or column.
That is a very good point and something we are evaluating at the moment.

The highlight with Blackbird is great to see but we were confused by the intention of the Github fork. The codebase suggests that it is more than just a proof of concept to highlight BioJS as there is project-specific branding. Would the authors consider clarifying their position with Blackbird?
After feedback from Holly Bik (Principal Investigator on the Phinch framework) we agreed to remove the Blackbird branding and instead merge our improvements back into Phinch. Therefore, we removed references to Blackbird from the manuscript. For more details see the referee report by Holly Bik (18 Oct 2016) and this discussion on GitHub:
https://github.com/PitchInteractiveInc/Phinch/issues/63

The primary motivator for the development of BIOM-format 2.1.0 were scaling limitations inherent with the JSON-based representation of 1.0.0. Specifically, the "data" key of the JSON string must be parsed in full in order to random access to individual sample or observation data. This removes the possibility of algorithms which depend on efficient random access patterns for data too large for main memory. Additionally, the overhead associated with representing a large JSON object in memory is high. While we acknowledge HDF5 possesses challenges for web-based interaction with these data, it is important to note that the 1.0.0 JSON-based format is not recommended for modern sized studies using hundreds to thousands to tens of thousands of samples.
This is a valid point. By using the JSON representation for our library we re-introduce the limitations of BIOM-format 1.0. We hope to support the HDF5 format in the future. However even with support of HDF5 loading full tables with tens of thousands of samples into the browser might be too memory intensive. Therefore, the next thing we would like to try is the extension of the conversion server with the communication API as you suggested. We added a short paragraph clearly stating our shortcoming and discussing the possible solution at the end of the Application section.

The use of the conversion server is very cool and could be taken a step further by layering a light communication API on top to allow a client to request arbitrary samples. This

separation would remove the burden of the client needing to read HDF5 formatted files, greatly lower the memory footprint of the client, and likely be more performant than a pure client-side model as the client would only need to know about what it had requested. This expansion of biojs-io-biom, in our opinion, would have the greatest impact for expanding the use of BIOM formatted data within a web application.
This is a great suggestion and we are eager to work on that for the next major release. We also added this as a future prospect to the manuscript.

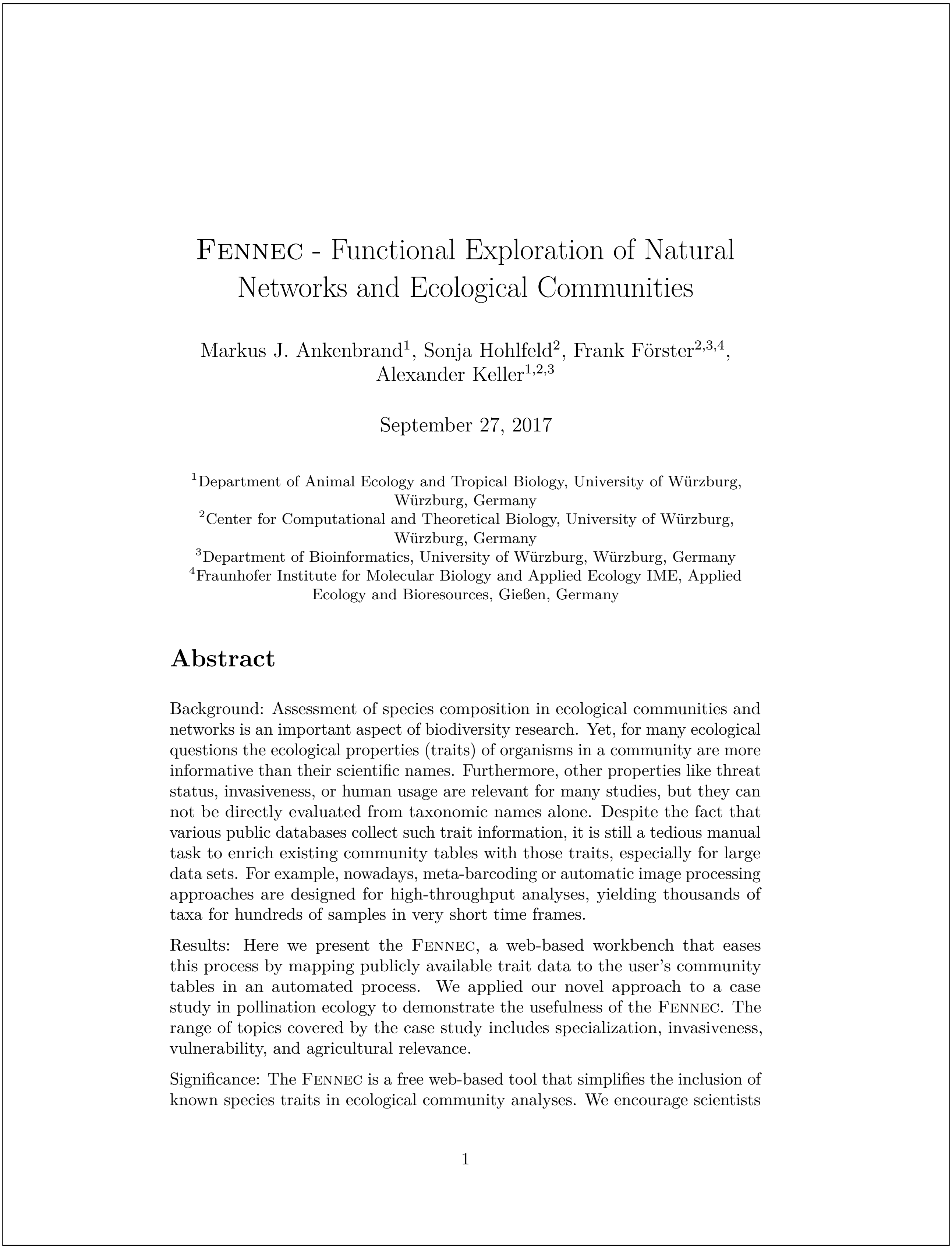
When the authors refer to BIOM v2, we believe they are actually referring to BIOM v2.1.0. There are important distinctions between the format versions. Would the authors consider clarifying the minor version number in discussion?
We added the minor version number whenever we refer to the BIOM format. We left the patch level out as the documentation on biom-format.org only lists the three versions (1.0, 2.0, 2.1). If you feel that the patch level is relevant as well we will gladly add that, too.

The two uses of "accession functions" reads awkwardly as these types of methods are generally described as "accessor functions." Would the authors consider revising the phrasing?
Thanks a lot. We revised the phrasing.

Competing Interests: No competing interests were disclosed.

5.7 fennec - functional exploration of natural networks and ecological communities

– published in *bioRχiv* –



Fennec - Functional Exploration of Natural Networks and Ecological Communities

Markus J. Ankenbrand[1], Sonja Hohlfeld[2], Frank Förster[2,3,4], Alexander Keller[1,2,3]

September 27, 2017

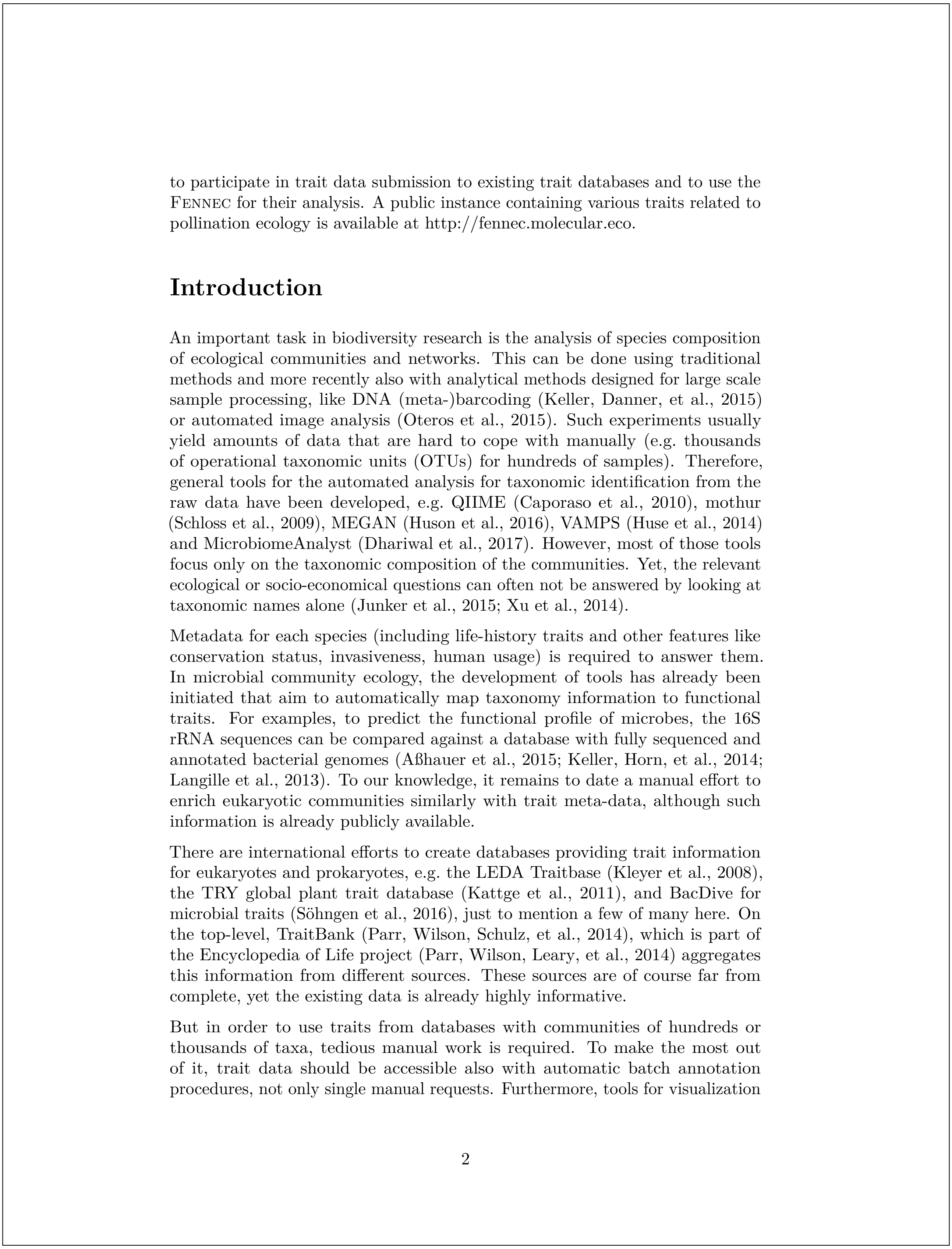
[1]Department of Animal Ecology and Tropical Biology, University of Würzburg, Würzburg, Germany
[2]Center for Computational and Theoretical Biology, University of Würzburg, Würzburg, Germany
[3]Department of Bioinformatics, University of Würzburg, Würzburg, Germany
[4]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Applied Ecology and Bioresources, Gießen, Germany

Abstract

Background: Assessment of species composition in ecological communities and networks is an important aspect of biodiversity research. Yet, for many ecological questions the ecological properties (traits) of organisms in a community are more informative than their scientific names. Furthermore, other properties like threat status, invasiveness, or human usage are relevant for many studies, but they can not be directly evaluated from taxonomic names alone. Despite the fact that various public databases collect such trait information, it is still a tedious manual task to enrich existing community tables with those traits, especially for large data sets. For example, nowadays, meta-barcoding or automatic image processing approaches are designed for high-throughput analyses, yielding thousands of taxa for hundreds of samples in very short time frames.

Results: Here we present the Fennec, a web-based workbench that eases this process by mapping publicly available trait data to the user's community tables in an automated process. We applied our novel approach to a case study in pollination ecology to demonstrate the usefulness of the Fennec. The range of topics covered by the case study includes specialization, invasiveness, vulnerability, and agricultural relevance.

Significance: The Fennec is a free web-based tool that simplifies the inclusion of known species traits in ecological community analyses. We encourage scientists



to participate in trait data submission to existing trait databases and to use the Fennec for their analysis. A public instance containing various traits related to pollination ecology is available at http://fennec.molecular.eco.

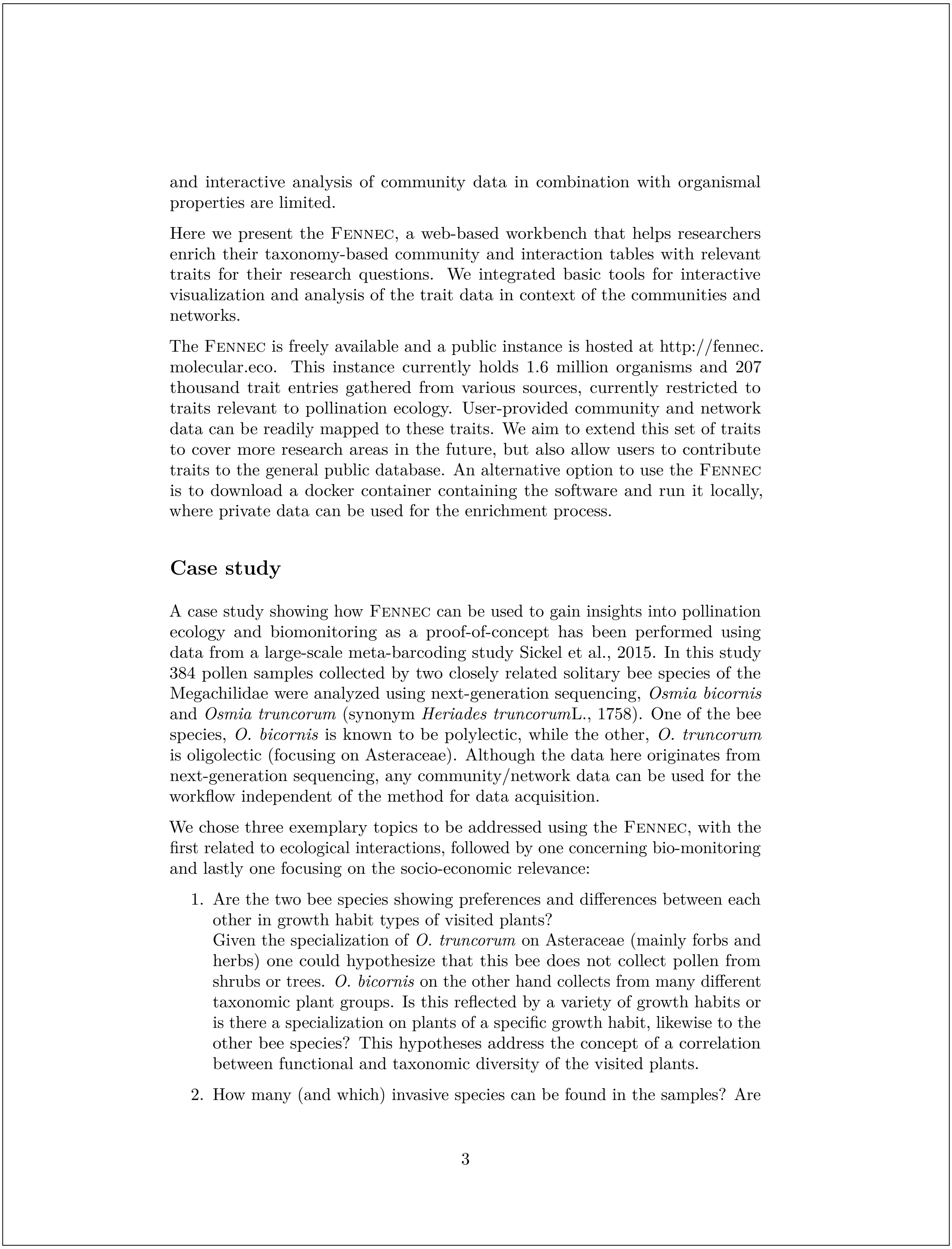
Introduction

An important task in biodiversity research is the analysis of species composition of ecological communities and networks. This can be done using traditional methods and more recently also with analytical methods designed for large scale sample processing, like DNA (meta-)barcoding (Keller, Danner, et al., 2015) or automated image analysis (Oteros et al., 2015). Such experiments usually yield amounts of data that are hard to cope with manually (e.g. thousands of operational taxonomic units (OTUs) for hundreds of samples). Therefore, general tools for the automated analysis for taxonomic identification from the raw data have been developed, e.g. QIIME (Caporaso et al., 2010), mothur (Schloss et al., 2009), MEGAN (Huson et al., 2016), VAMPS (Huse et al., 2014) and MicrobiomeAnalyst (Dhariwal et al., 2017). However, most of those tools focus only on the taxonomic composition of the communities. Yet, the relevant ecological or socio-economical questions can often not be answered by looking at taxonomic names alone (Junker et al., 2015; Xu et al., 2014).

Metadata for each species (including life-history traits and other features like conservation status, invasiveness, human usage) is required to answer them. In microbial community ecology, the development of tools has already been initiated that aim to automatically map taxonomy information to functional traits. For examples, to predict the functional profile of microbes, the 16S rRNA sequences can be compared against a database with fully sequenced and annotated bacterial genomes (Aßhauer et al., 2015; Keller, Horn, et al., 2014; Langille et al., 2013). To our knowledge, it remains to date a manual effort to enrich eukaryotic communities similarly with trait meta-data, although such information is already publicly available.

There are international efforts to create databases providing trait information for eukaryotes and prokaryotes, e.g. the LEDA Traitbase (Kleyer et al., 2008), the TRY global plant trait database (Kattge et al., 2011), and BacDive for microbial traits (Söhngen et al., 2016), just to mention a few of many here. On the top-level, TraitBank (Parr, Wilson, Schulz, et al., 2014), which is part of the Encyclopedia of Life project (Parr, Wilson, Leary, et al., 2014) aggregates this information from different sources. These sources are of course far from complete, yet the existing data is already highly informative.

But in order to use traits from databases with communities of hundreds or thousands of taxa, tedious manual work is required. To make the most out of it, trait data should be accessible also with automatic batch annotation procedures, not only single manual requests. Furthermore, tools for visualization



and interactive analysis of community data in combination with organismal properties are limited.

Here we present the Fennec, a web-based workbench that helps researchers enrich their taxonomy-based community and interaction tables with relevant traits for their research questions. We integrated basic tools for interactive visualization and analysis of the trait data in context of the communities and networks.

The Fennec is freely available and a public instance is hosted at http://fennec.molecular.eco. This instance currently holds 1.6 million organisms and 207 thousand trait entries gathered from various sources, currently restricted to traits relevant to pollination ecology. User-provided community and network data can be readily mapped to these traits. We aim to extend this set of traits to cover more research areas in the future, but also allow users to contribute traits to the general public database. An alternative option to use the Fennec is to download a docker container containing the software and run it locally, where private data can be used for the enrichment process.

Case study

A case study showing how Fennec can be used to gain insights into pollination ecology and biomonitoring as a proof-of-concept has been performed using data from a large-scale meta-barcoding study Sickel et al., 2015. In this study 384 pollen samples collected by two closely related solitary bee species of the Megachilidae were analyzed using next-generation sequencing, *Osmia bicornis* and *Osmia truncorum* (synonym *Heriades truncorum*L., 1758). One of the bee species, *O. bicornis* is known to be polylectic, while the other, *O. truncorum* is oligolectic (focusing on Asteraceae). Although the data here originates from next-generation sequencing, any community/network data can be used for the workflow independent of the method for data acquisition.

We chose three exemplary topics to be addressed using the Fennec, with the first related to ecological interactions, followed by one concerning bio-monitoring and lastly one focusing on the socio-economic relevance:

1. Are the two bee species showing preferences and differences between each other in growth habit types of visited plants?
   Given the specialization of *O. truncorum* on Asteraceae (mainly forbs and herbs) one could hypothesize that this bee does not collect pollen from shrubs or trees. *O. bicornis* on the other hand collects from many different taxonomic plant groups. Is this reflected by a variety of growth habits or is there a specialization on plants of a specific growth habit, likewise to the other bee species? This hypotheses address the concept of a correlation between functional and taxonomic diversity of the visited plants.

2. How many (and which) invasive species can be found in the samples? Are

there vulnerable species in the samples? Is the amount of invasive and vulnerable species visited similar in all of the samples and by both species? Monitoring the ranges of invasive as well as threatened plant species is an important task in conservation (Darling et al., 2007; Stout, Jane C. et al., 2009). Using pollen data collected by bees, presence of both types can be monitored by mapping conservation relevant traits to the network data. Further, pollination services by the bee species to both types can be identified.

3. Which plants visited by the bees show agricultural relevance to humans and what is their relative amount compared to the remaining plant species? Bees are used commercially to provide pollination services to agriculturally relevant plants (Klatt et al., 2013). Using traits as *agricultural usage* allow to identify how specific the respective bees were in visiting such plants. On the other hand, solitary bees are important agents to ensure the pollination of wild plant species (Garibaldi et al., 2013), and using these traits it can be monitored whether the bees are mainly attracted to mass flowering crops or also visit other plants in agriculturally shaped landscapes.

Material and Methods

Code Implementation and Accessibility

The Fennec is a web application developed in PHP (http://php.net/) using the Symfony framework (https://symfony.com/) with a front-end using JavaScript (ES6) for interactivity. Server side functionality is bundled in modular web services that are called from the front-end via AJAX requests. Layout and interactivity are provided by multiple well established libraries including bootstrap (https://getbootstrap.com/), jQuery-ui (https://jqueryui.com/), react (https://facebook.github.io/react/), lodash (https://lodash.com/), datatables (https://datatables.net/), and plotly.js (https://plot.ly/javascript/). The code quality is ensured using unit tests and strategies for continuous integration. All data is stored in a PostgreSQL (https://www.postgresql.org/) database, which includes taxonomic, trait, citation and further meta-information (Suppl. Fig. 1). Database accession is handled via the doctrine object relational mapper (http://www.doctrine-project.org/). The community and network data provided by users are uploaded and stored in BIOM format (version 1.0) (McDonald et al., 2012) using the biojs-io-biom library (Ankenbrand et al., 2017).

There are three ways to use the Fennec workbench:

1. Public Instance: We have set up a public instance of the Fennec available at http://fennec.molecular.eco. It's database currently hosts trait data related to pollination ecology and is gathered from various sources. This dataset includes also all traits of the case study presented here, yet not

exclusively. The database is subject to constant further extension with more traits, yet our main goal is to maintain high quality of the data available here. User-provided network and community data is private per default, requiring the user to authenticate using a Fennec, GitHub or Google account.

2. Local Instances: All program code is open-source (MIT License) and freely available at the public repository of GitHub: https://github.com/molbiodiv/fennec. Alongside the pure code, there are ready-to-use docker containers available, which include pre-configured instances to be run in a virtualization environment. More information about the docker environment can be found under https://www.docker.com. These local instances can be set up to be accessible only in a specific local network, so that the software can be run in a restrictive way by workgroups or users. These databases can be filled directly with arbitrary trait data not limited to that included in the public instance.
3. Application Programming Interface (API): We also provide an open API that allows third-party programs to make calls to the public instance, or if available also local instances. We currently implemented an R package that makes use of this API. The usage is not limited to this and we encourage software developers to use this API for integration.

Extensive documentation on the code, but also tutorials for users and guides for administrators to host local instances and software developers to use the API are available at both, the GitHub repository and the public instance.

Mapping community and network data to traits

Minimum requirement for using the Fennec is to provide a community or network table, including taxa as rows and samples (communities)/ taxa (networks) as columns. The cells are considered to represent abundances for the respective combinations, but also presence/absence data can be used. Beyond this table, users may also provide own taxonomy data for taxa (to use an alternative to our default NCBI taxonomy (Federhen, 2012)) and meta-data for the samples. These tables may be uploaded separately as tab-separated text files, or combined in BIOM format (Ankenbrand et al., 2017; McDonald et al., 2012) (figure 1). All tables can be managed using the project page of Fennec (figure 3B). Depending on the user input, taxa are mapped using their scientific names or database identifiers like NCBI-taxonomy-ID (Federhen, 2012) or EOL-ID (Parr, Wilson, Leary, et al., 2014).

Traits to be analyzed with this data can be explored and selected via the intuitive web interface, and added as meta-data to the project (figure 3A). All trait data available for this trait and the taxa of interest are automatically linked into the dataset. If multiple values are available for a single trait and organism combination, they are automatically aggregated, i.e. categorical traits are

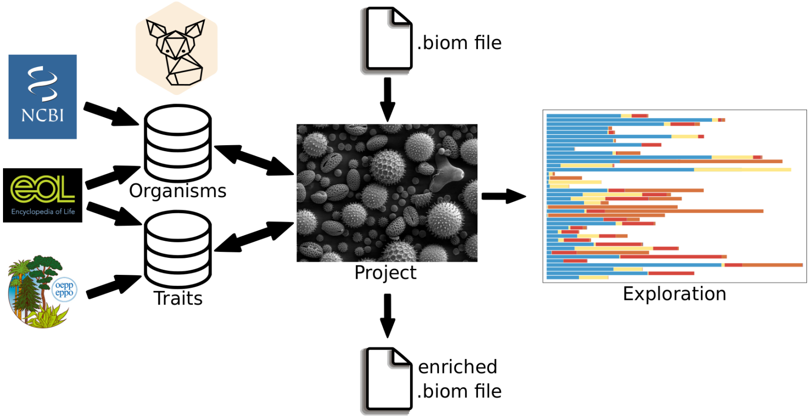
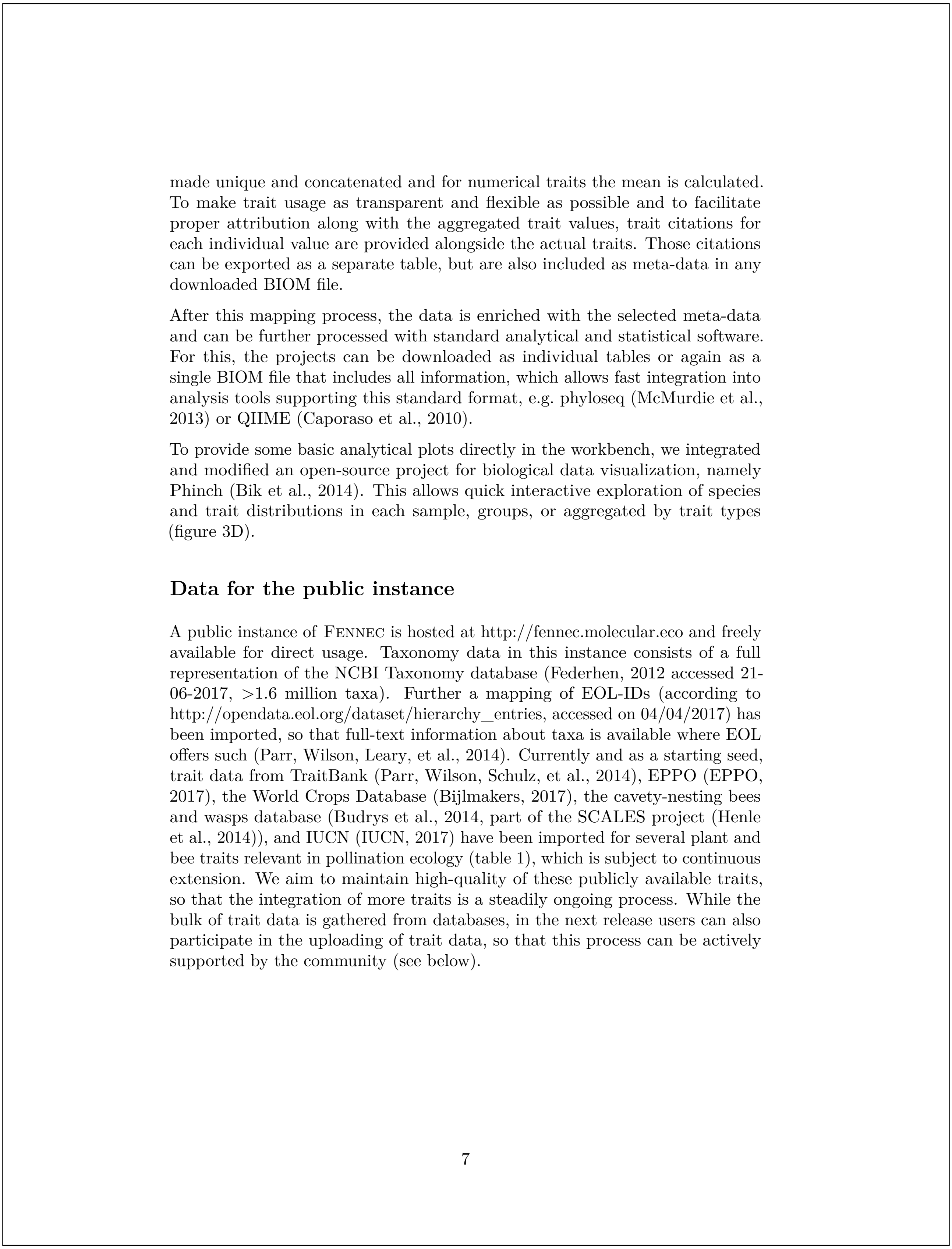


Figure 1: General structure of Fennec. Organisms and traits from different sources like NCBI (Coordinators, 2017), EOL (Parr, Wilson, Leary, et al., 2014), and EPPO are stored in Fennecs database by the administrator. The user imports a community project e.g. in biom format. The organisms in the community are mapped against those in the database and associated traits are used to enrich the metadata. The trait composition can be interactively explored and enriched projects exported e.g. in biom format.



made unique and concatenated and for numerical traits the mean is calculated. To make trait usage as transparent and flexible as possible and to facilitate proper attribution along with the aggregated trait values, trait citations for each individual value are provided alongside the actual traits. Those citations can be exported as a separate table, but are also included as meta-data in any downloaded BIOM file.

After this mapping process, the data is enriched with the selected meta-data and can be further processed with standard analytical and statistical software. For this, the projects can be downloaded as individual tables or again as a single BIOM file that includes all information, which allows fast integration into analysis tools supporting this standard format, e.g. phyloseq (McMurdie et al., 2013) or QIIME (Caporaso et al., 2010).

To provide some basic analytical plots directly in the workbench, we integrated and modified an open-source project for biological data visualization, namely Phinch (Bik et al., 2014). This allows quick interactive exploration of species and trait distributions in each sample, groups, or aggregated by trait types (figure 3D).

Data for the public instance

A public instance of Fennec is hosted at http://fennec.molecular.eco and freely available for direct usage. Taxonomy data in this instance consists of a full representation of the NCBI Taxonomy database (Federhen, 2012 accessed 21-06-2017, >1.6 million taxa). Further a mapping of EOL-IDs (according to http://opendata.eol.org/dataset/hierarchy_entries, accessed on 04/04/2017) has been imported, so that full-text information about taxa is available where EOL offers such (Parr, Wilson, Leary, et al., 2014). Currently and as a starting seed, trait data from TraitBank (Parr, Wilson, Schulz, et al., 2014), EPPO (EPPO, 2017), the World Crops Database (Bijlmakers, 2017), the cavety-nesting bees and wasps database (Budrys et al., 2014, part of the SCALES project (Henle et al., 2014)), and IUCN (IUCN, 2017) have been imported for several plant and bee traits relevant in pollination ecology (table 1), which is subject to continuous extension. We aim to maintain high-quality of these publicly available traits, so that the integration of more traits is a steadily ongoing process. While the bulk of trait data is gathered from databases, in the next release users can also participate in the uploading of trait data, so that this process can be actively supported by the community (see below).

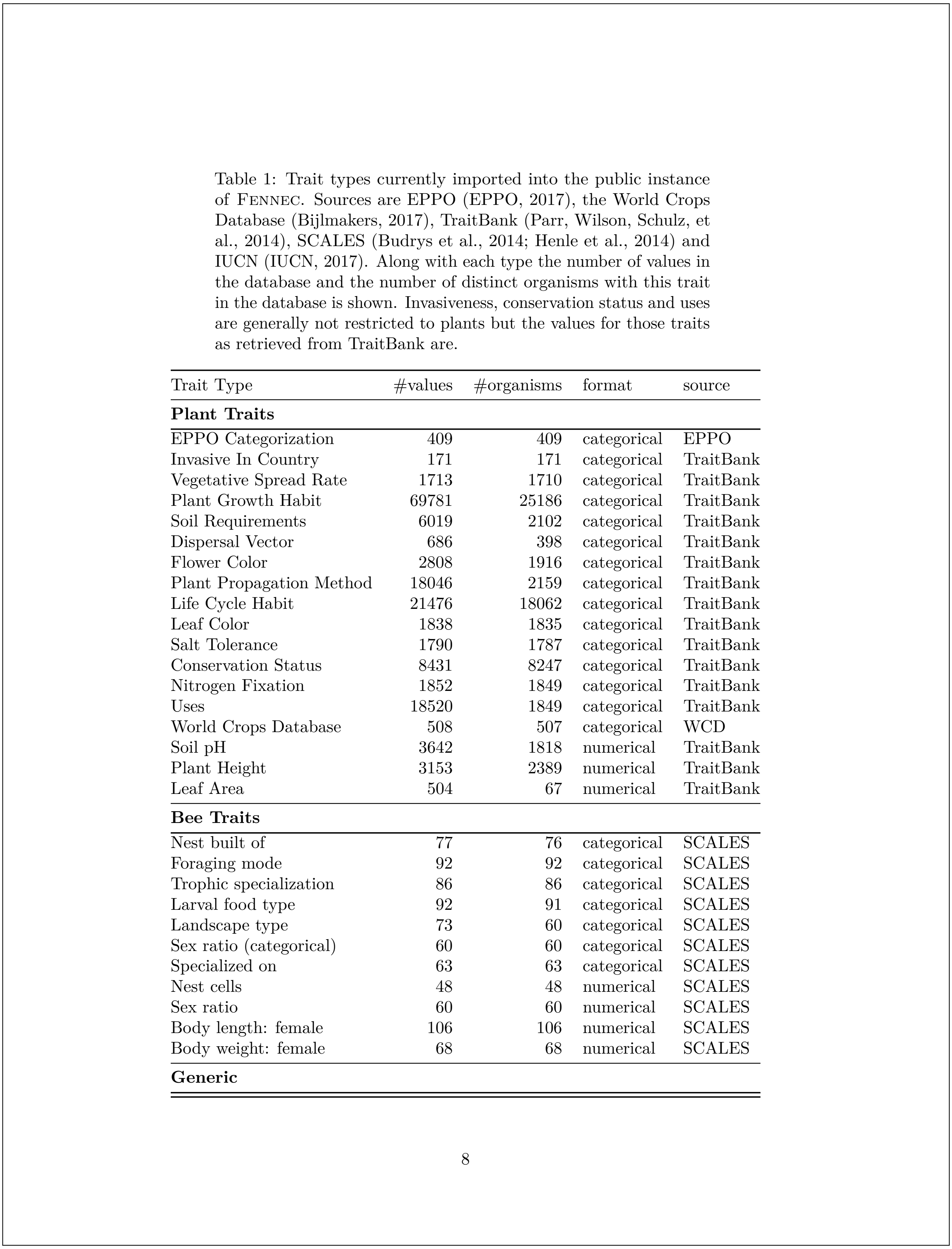


Table 1: Trait types currently imported into the public instance of Fennec. Sources are EPPO (EPPO, 2017), the World Crops Database (Bijlmakers, 2017), TraitBank (Parr, Wilson, Schulz, et al., 2014), SCALES (Budrys et al., 2014; Henle et al., 2014) and IUCN (IUCN, 2017). Along with each type the number of values in the database and the number of distinct organisms with this trait in the database is shown. Invasiveness, conservation status and uses are generally not restricted to plants but the values for those traits as retrieved from TraitBank are.

Trait Type	#values	#organisms	format	source
Plant Traits				
EPPO Categorization	409	409	categorical	EPPO
Invasive In Country	171	171	categorical	TraitBank
Vegetative Spread Rate	1713	1710	categorical	TraitBank
Plant Growth Habit	69781	25186	categorical	TraitBank
Soil Requirements	6019	2102	categorical	TraitBank
Dispersal Vector	686	398	categorical	TraitBank
Flower Color	2808	1916	categorical	TraitBank
Plant Propagation Method	18046	2159	categorical	TraitBank
Life Cycle Habit	21476	18062	categorical	TraitBank
Leaf Color	1838	1835	categorical	TraitBank
Salt Tolerance	1790	1787	categorical	TraitBank
Conservation Status	8431	8247	categorical	TraitBank
Nitrogen Fixation	1852	1849	categorical	TraitBank
Uses	18520	1849	categorical	TraitBank
World Crops Database	508	507	categorical	WCD
Soil pH	3642	1818	numerical	TraitBank
Plant Height	3153	2389	numerical	TraitBank
Leaf Area	504	67	numerical	TraitBank
Bee Traits				
Nest built of	77	76	categorical	SCALES
Foraging mode	92	92	categorical	SCALES
Trophic specialization	86	86	categorical	SCALES
Larval food type	92	91	categorical	SCALES
Landscape type	73	60	categorical	SCALES
Sex ratio (categorical)	60	60	categorical	SCALES
Specialized on	63	63	categorical	SCALES
Nest cells	48	48	numerical	SCALES
Sex ratio	60	60	numerical	SCALES
Body length: female	106	106	numerical	SCALES
Body weight: female	68	68	numerical	SCALES
Generic				

Trait Type	#values	#organisms	format	source
IUCN Red List	44677	44469	categorical	IUCN

Importing organism and trait data

For both, public and local instances, traits are imported using a simple table format containing the organism and trait value, a citation as well as optionally an ontology URL, and source URL as columns, and each entry as a new row. This allows for easy import of traits from various sources. A template for the format for upload is available alongside the Fennec GitHub repository or the public instance.

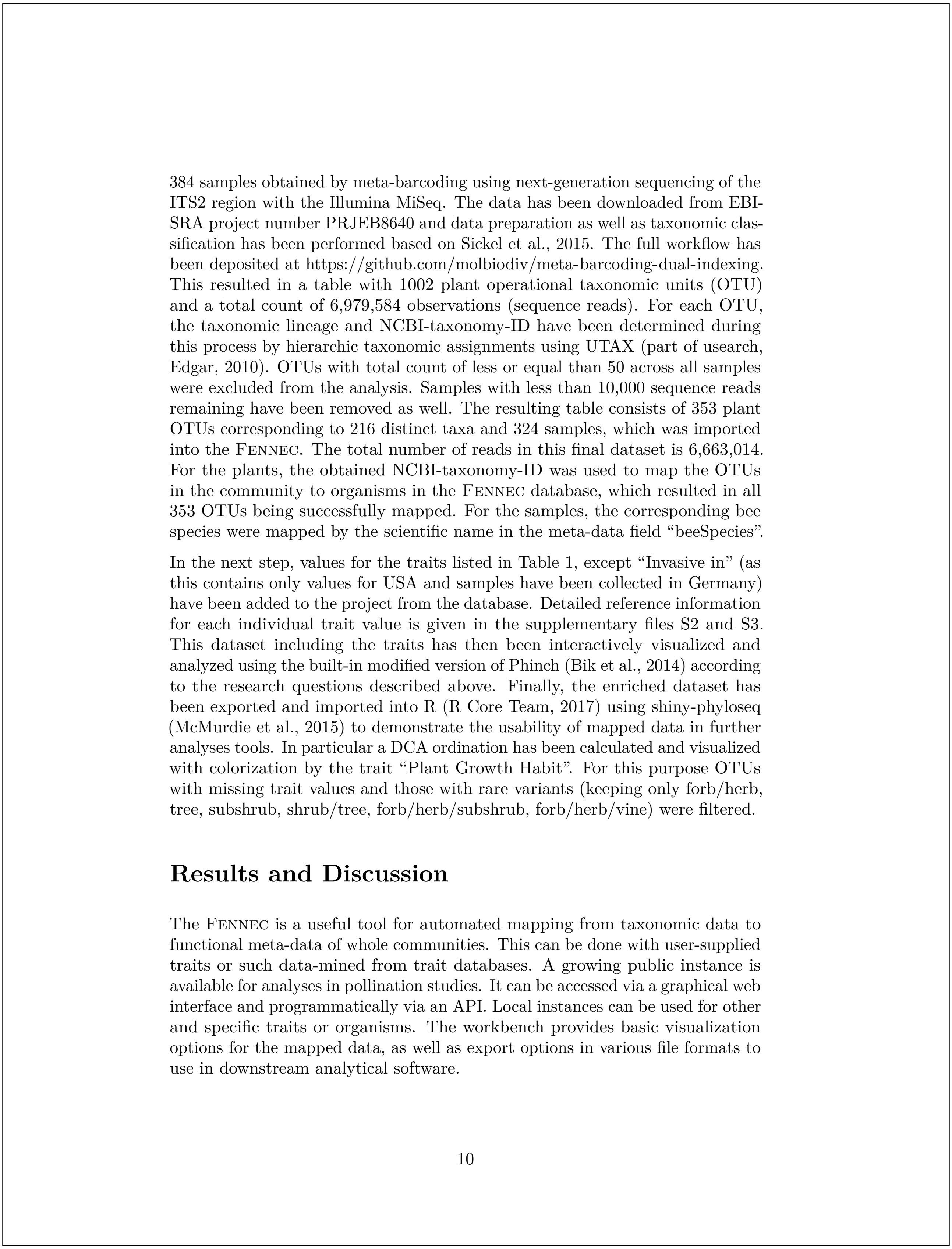
Two trait formats are currently supported: categorical and numerical. Categorical traits may also include an ontology URL for their value, supporting the hierarchical classification characteristics of some traits. Numerical trait types may be uploaded with an associated unit.

Currently users can upload their own traits as project specific meta-data. It is planned for future releases to upload traits to the public instance using a form. After the submission and verification process, the user may work on this meta-data privately or choose to make it publicly accessible for other researchers. All data marked as public will then undergo a limited manual verification, which e.g. ensures that units are correctly standardized, but does not verify the correctness of the underlying data. Therefore, the user-name of the uploader will be permanently linked to the data to be able to address future changes and updates, and it is required to provide citation information for public records.

For local instances, arbitrary organism and trait data can be imported into the Fennec database by the local administrator using the command line interface (CLI), which are then not subject to a central verification process, but available instantly. This allows for creation of instances tailored to specific organism groups and associated research questions, with the responsibility of the administrator to ensure quality of the imported data. For these public instances, either the same NCBI-taxonomy data can be used or custom taxonomy data provided. Each imported organism receives a unique Fennec-ID which can be linked on-the-fly to other identifiers like NCBI-Taxonomy-IDs or EOL-IDs. The linked EOL-ID is used to provide dynamic content for each organism using the EOL application programming interface (API), where EOL provides such (figure 3C).

Case Study

To demonstrate the analytical potential of the Fennec, we use it to analyse data obtained from pollen collections of the two solitary megachilid bees *Osmia bicornis* and *O. truncorum* in Germany (Sickel et al., 2015). The dataset consists of



384 samples obtained by meta-barcoding using next-generation sequencing of the ITS2 region with the Illumina MiSeq. The data has been downloaded from EBI-SRA project number PRJEB8640 and data preparation as well as taxonomic classification has been performed based on Sickel et al., 2015. The full workflow has been deposited at https://github.com/molbiodiv/meta-barcoding-dual-indexing. This resulted in a table with 1002 plant operational taxonomic units (OTU) and a total count of 6,979,584 observations (sequence reads). For each OTU, the taxonomic lineage and NCBI-taxonomy-ID have been determined during this process by hierarchic taxonomic assignments using UTAX (part of usearch, Edgar, 2010). OTUs with total count of less or equal than 50 across all samples were excluded from the analysis. Samples with less than 10,000 sequence reads remaining have been removed as well. The resulting table consists of 353 plant OTUs corresponding to 216 distinct taxa and 324 samples, which was imported into the Fennec. The total number of reads in this final dataset is 6,663,014. For the plants, the obtained NCBI-taxonomy-ID was used to map the OTUs in the community to organisms in the Fennec database, which resulted in all 353 OTUs being successfully mapped. For the samples, the corresponding bee species were mapped by the scientific name in the meta-data field “beeSpecies”.

In the next step, values for the traits listed in Table 1, except “Invasive in” (as this contains only values for USA and samples have been collected in Germany) have been added to the project from the database. Detailed reference information for each individual trait value is given in the supplementary files S2 and S3. This dataset including the traits has then been interactively visualized and analyzed using the built-in modified version of Phinch (Bik et al., 2014) according to the research questions described above. Finally, the enriched dataset has been exported and imported into R (R Core Team, 2017) using shiny-phyloseq (McMurdie et al., 2015) to demonstrate the usability of mapped data in further analyses tools. In particular a DCA ordination has been calculated and visualized with colorization by the trait “Plant Growth Habit”. For this purpose OTUs with missing trait values and those with rare variants (keeping only forb/herb, tree, subshrub, shrub/tree, forb/herb/subshrub, forb/herb/vine) were filtered.

Results and Discussion

The Fennec is a useful tool for automated mapping from taxonomic data to functional meta-data of whole communities. This can be done with user-supplied traits or such data-mined from trait databases. A growing public instance is available for analyses in pollination studies. It can be accessed via a graphical web interface and programmatically via an API. Local instances can be used for other and specific traits or organisms. The workbench provides basic visualization options for the mapped data, as well as export options in various file formats to use in downstream analytical software.

Case study

To show the potential of the Fennec to be used in ecological analysis, we conducted a case study as proof-of-concept for a pollen meta-barcoding data. We address multiple ecological questions and highlights some use cases, where automatic integration of public trait data with the FENNEC has been performed.

Are the two bee species showing preferences and differences between each other in growth habit types of visited plants?

A breakdown of the trait "Plant Growth Habit" for the two bee species separately (visualized via "Donut Partition Chart") reveals that for *O. truncorum* 89% of the taxonomic assignments were mappable to the trait, which resulted in absolute a dominance of "forb/herb" with 87%. This matches our expectations as this bee is specialized on Asteraceae which mostly show this habit.

For *O. bicornis*, 95% of the sequence data was assignable to "Plant Growth Habit", also with "forb/herb" with 65% being the most abundant, but a still considerable amount of 24% as "tree". Likewise to taxonomic specialization, no indication for a specialization on a specific plant growth habit is apparent.

Another interesting observation is the trait coverage when taking abundance into account. While only 85% of OTUs have a value for "Plant Growth Habit", those OTUs contribute 93% to the entire community. Thus the OTUs with missing trait are relatively rare in the community, with the more abundant ones being well-studied in terms of trait data.

Automatically mapped trait data also helps in interpretation of beta-diversity turnover between samples, here collected pollens. For example, ordinations can be visualized with trait data, in our case "Plant Growth Habit", as a split-plot with samples shaped by bee species and plant taxa colored by Plant Growth Habit (figure 2). In our case study, samples are separated as expected by bee species on the first ordination axis with all samples from *O. truncorum* mostly isolated on the right hand side. OTUs localized similarly with possible values for ordination axis 1 were almost exclusively forbs and herbs. The variation of this bee species on the second axis is negligible. For *O. bicornis* there is a substantial spread particularly on the second axis, where plants of type tree seem to concentrate in the upper part. The trait data helps to understand the ecology behind the dataset, indicating plant turnover and eventually also location and landscape changes to be represented on the second axis.

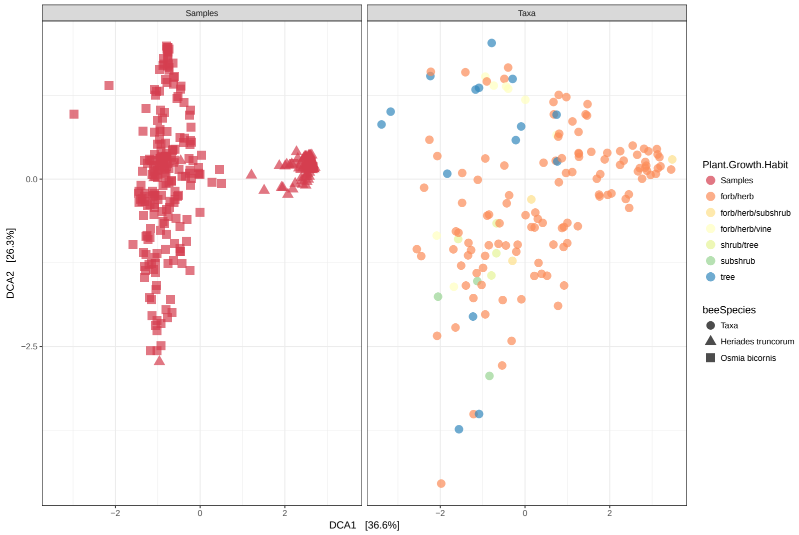


Figure 2: Splitplot of a DCA ordination. Samples are in the left facet with shapes according to by bee species. OTUs are in the right facet with points colored by growth habit (filtered for most common growth habits, species with missing trait have been removed). Samples split nicely by bee on the first axis with *O. truncorum* on the right hand side. The OTUs on the right hand side of the ordination are as expected mainly forb/herb. For *O. bicornis* there is a substantial spread on the second axis.

How many (and which) invasive species can be found in the samples? Are there vulnerable species in the samples? Is the amount of invasive and vulnerable species visited similar in all of the samples and by both species?

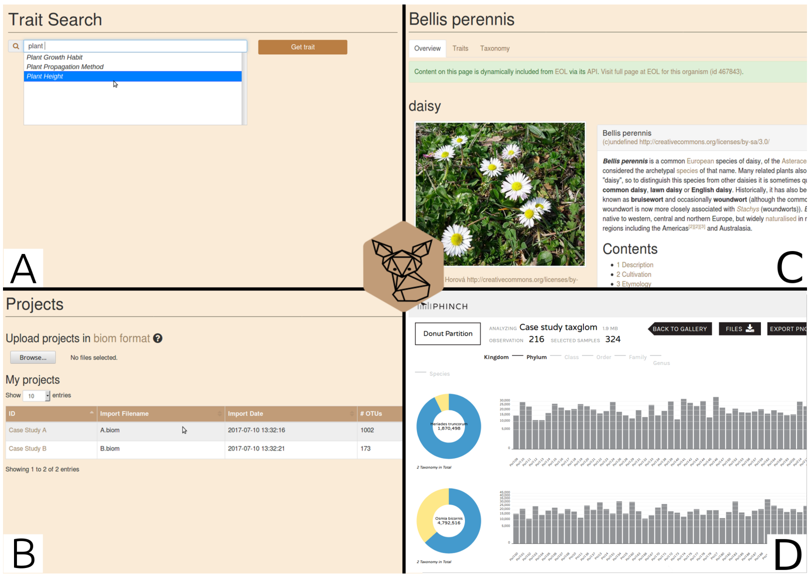
The trait "EPPO Categorization" was mapped to our pollen collection data to determine if and to what extend the samples contain species that are regarded as invasive in Europe. One of the visualization methods of the Phinch suite that is integrated into the Fennec, the "Bubble Chart", has been applied to explore this trait. It reveals three samples containing high numbers of invasive species (PoJ74, PoJ236, PoJ244). Further inspection with the integrated meta-data tables showed that PoJ74 and PoJ244 have more than 1000 counts of *Solidago canadensis*, each while PoJ236 has a count of 2779 for *Helianthus tuberosus*.

The trait data is thus sortable regarding abundance of specific traits, referable to organism information but also to samples and their corresponding geographical locations if the data has been collected in such way. It might thus serve as indicator for occurrence of invasive species in geographic regions and used to monitor the spread of invasive species over space and time.

Regarding the occurrence of species with respect to threat status, the pollen data was automatically mapped to the IUCN red list data and the distribution of vulnerable species (as listed by the IUCN) across samples was visualized using the "Bubble Chart", but also a "Taxonomy Bar Chart". These charts illustrate that multiple samples consist almost entirely of "near threatened" species, particularly *Juglans regia*, the english walnut, which experienced strong declines through anthropogenic overuse and lack of replacement plantings. As indicated by the data, it served as a major nutrient source for individual investigated bees.

Which plants visited by the bees are agriculturally relevant to humans and what is their relative amount compared to the remaining plant species?

Finally ecologists (especially in the field of conservation) are often in the difficult situation to somehow quantify economic value of ecosystem services like pollination (Hanley et al., 2015). The Fennec helps in addressing such socio-economic questions by including human usage (as crop) as a trait. All plants listed in the World Crops Database are known to be cultivated by humans for specific purposes (Bijlmakers, 2017). The "Donut Partition Chart" for this trait reveals that 36.7% of plants collected by *O. bicornis* and 7.3% of plants collected by *O. truncorum* are listed in that database (figure 4). This does not yet give more information like the category of crop (e.g. fruits, vegetables, nuts, wood product, etc.) or a real monetary quantification. However this is not a limitation of Fennec but of the underlying data (i.e. if this data is available it can be imported into Fennec and is then automatically available for the community of interest).



Content on this page is dynamically included from EOL via its API. Visit full page at EOL for this organism (id 467843).

Bellis perennis
(c)undefined http://creativecommons.org/licenses/by-sa/3.0/

Bellis perennis is a common European species of daisy, of the Asterace
considered the archetypal species of that name. Many related plants also
"daisy", so to distinguish this species from other daisies it is sometimes qu
common daisy, **lawn daisy** or **English daisy**. Historically, it has also be
known as **bruisewort** and occasionally **woundwort** (although the commo
woundwort is now more closely associated with *Stachys* (woundworts)). *B*
native to western, central and northern Europe, but widely naturalised in r
regions including the Americas[2][2][3] and Australasia.

Horová http://creativecommons.org/licenses/by-

Upload projects in biom format

No files selected.

ID	Import Filename	Import Date	# OTUs
Case Study A	A.biom	2017-07-10 13:32:16	1002
Case Study B	B.biom	2017-07-10 13:32:21	173

Showing 1 to 2 of 2 entries

ANALYZING Case study taxglom 1.9 MB
OBSERVATION 216 SELECTED SAMPLES 324

Figure 3: User interface of Fennec. A: Explore all traits which are stored in the database via the trait search. B: Upload community data tables, manage and analyze them. C: Get dynamic content for each organism using the API of EOL. D: Visualize data using the "Donut Partition Chart" of the Phinch suite that is integrated into the Fennec.

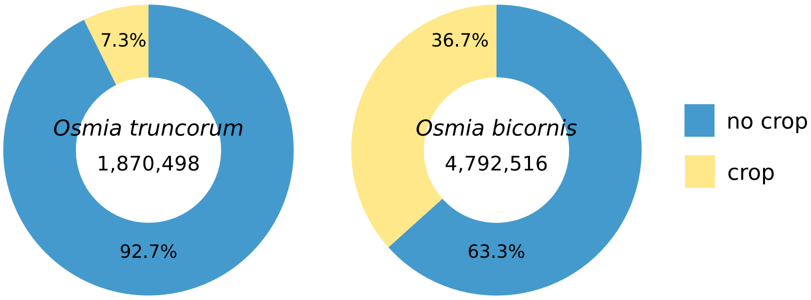


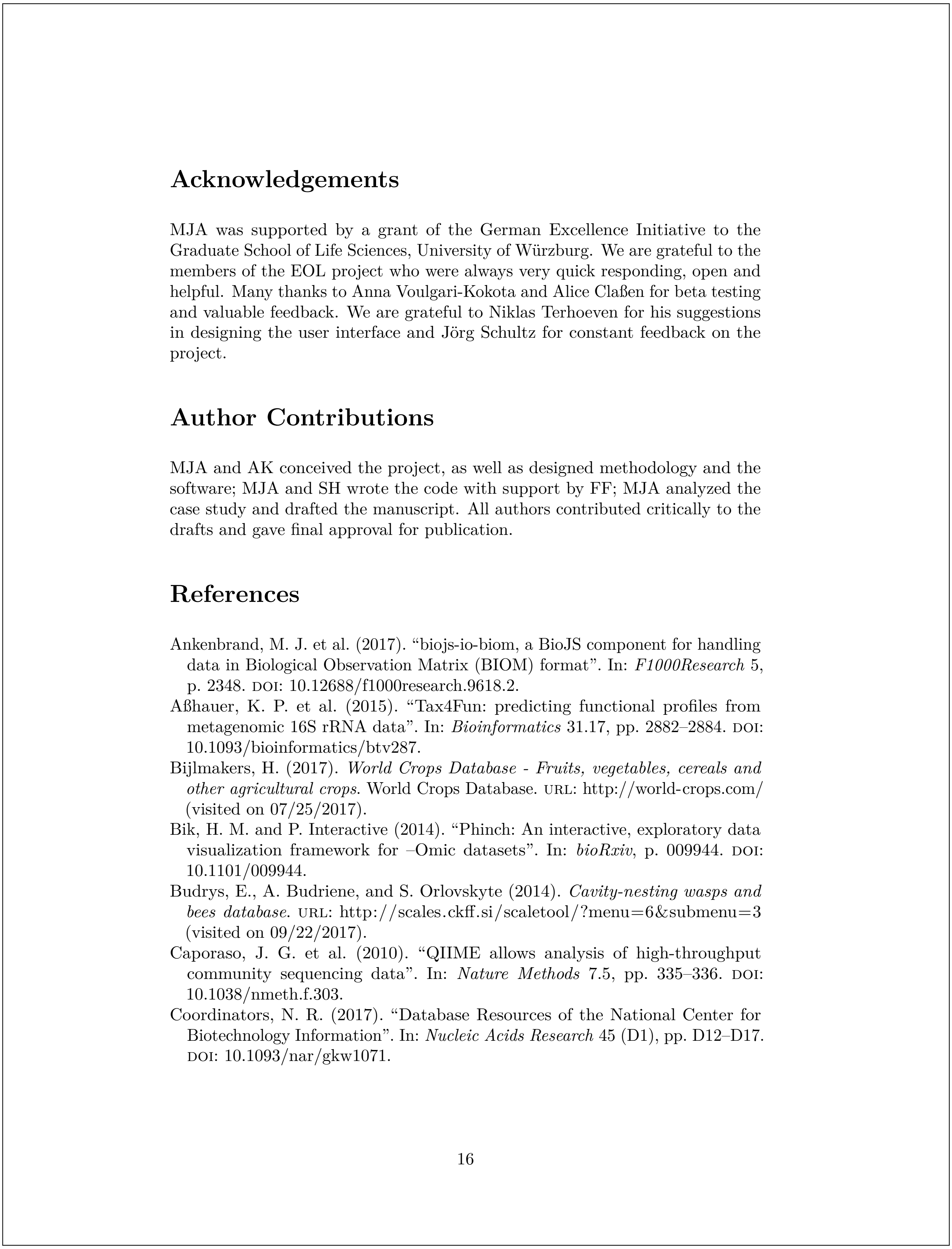
Figure 4: Partition donut charts for the trait "World Crops Database" separated by bee species. Plot has been created with the built in modified version of Phinch.

Outlook and limitations

Fennec has still a number of limitations that will be addressed in future releases. In particular common tasks like filtering and normalization have to be done prior to the upload, as that was not the main focus of the tool. Further, traits for bacteria have proven hard to use for standard 16S microbiome studies as classification can only go down to species or genus level (Werner et al., 2012), while traits may vary on the strain level (Truong et al., 2017). However, one of the main factors restricting Fennecs utility is the currently limited amount of trait data being available in a usable format. One thing that became apparent while building the Fennec is that a lot of trait data is available online but the majority can not be easily used because not adhering to the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Mons et al., 2017; Wilkinson et al., 2016). Licensing of the data is another common problem, data can only be efficiently re-used if it is open and citable in addition to being FAIR (Katz, 2017). An important step for trait data collectors to provide their data in this manner, is to guarantee that the data re-users are able to properly cite all data sources, e.g. by adhering to the FORCE11 data citation principles (Martone, 2014). Fennec supports this by preserving all relevant information. We therefore encourage trait data collectors to make their data available via existing platforms like TraitBank (Parr, Wilson, Schulz, et al., 2014), and with that also usable for downstream analysis tools like Fennec and ultimately to the whole research community.

Conclusion

Fennec as a tool provides valuable assistance to analyze ecological data in the context of organismal information. Both species traits and metadata like threat status and economical importance help to answer different kinds of questions. The public instance can be used as a reference, to try features of Fennec and analyze some datasets with data from other public databases. The possibility to host local instances with own data increases the range of applications. General problems are limited trait data availability, which is however increasing with time and the motivation of more and more scientific journals to make public data deposition mandatory for publications. Beside developing a public automatic mapping procedure, we also aim to demonstrate the importance to make trait data publicly available and how useful it can be in follow up studies. Thus encouraging scientists to submit their data to public databases. Despite those limitations we demonstrated that the Fennec is already able to facilitate ecological community analyses in the light of organismal information. It's usefulness is expected to increase due to continued development guided by user feedback, integration of more analysis tools, better taxonomic resolution, and increasing availability of suitable trait data.



Acknowledgements

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. We are grateful to the members of the EOL project who were always very quick responding, open and helpful. Many thanks to Anna Voulgari-Kokota and Alice Claßen for beta testing and valuable feedback. We are grateful to Niklas Terhoeven for his suggestions in designing the user interface and Jörg Schultz for constant feedback on the project.

Author Contributions

MJA and AK conceived the project, as well as designed methodology and the software; MJA and SH wrote the code with support by FF; MJA analyzed the case study and drafted the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Ankenbrand, M. J. et al. (2017). "biojs-io-biom, a BioJS component for handling data in Biological Observation Matrix (BIOM) format". In: *F1000Research* 5, p. 2348. doi: 10.12688/f1000research.9618.2.

Aßhauer, K. P. et al. (2015). "Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data". In: *Bioinformatics* 31.17, pp. 2882–2884. doi: 10.1093/bioinformatics/btv287.

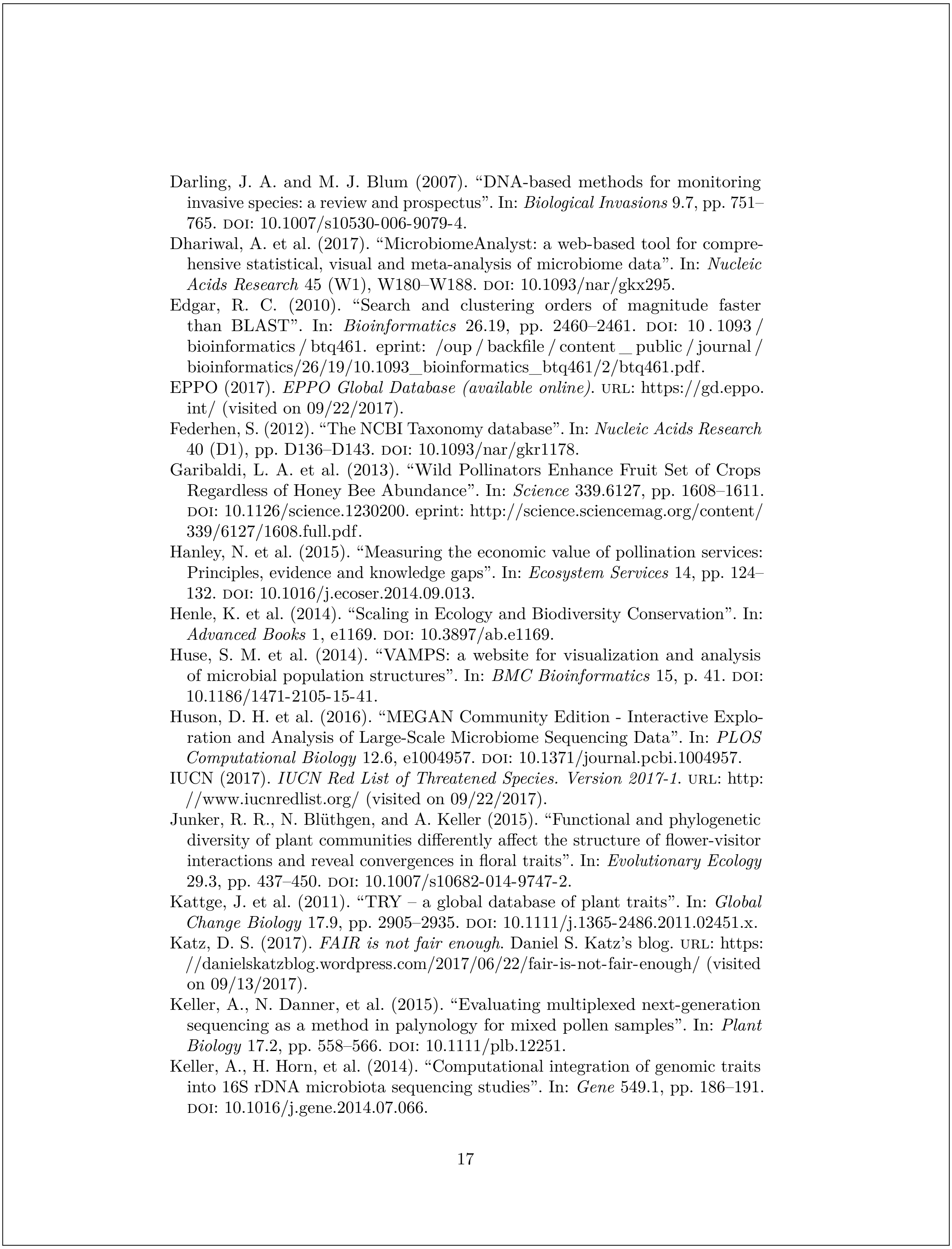
Bijlmakers, H. (2017). *World Crops Database - Fruits, vegetables, cereals and other agricultural crops.* World Crops Database. url: http://world-crops.com/ (visited on 07/25/2017).

Bik, H. M. and P. Interactive (2014). "Phinch: An interactive, exploratory data visualization framework for –Omic datasets". In: *bioRxiv*, p. 009944. doi: 10.1101/009944.

Budrys, E., A. Budriene, and S. Orlovskyte (2014). *Cavity-nesting wasps and bees database.* url: http://scales.ckff.si/scaletool/?menu=6&submenu=3 (visited on 09/22/2017).

Caporaso, J. G. et al. (2010). "QIIME allows analysis of high-throughput community sequencing data". In: *Nature Methods* 7.5, pp. 335–336. doi: 10.1038/nmeth.f.303.

Coordinators, N. R. (2017). "Database Resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 45 (D1), pp. D12–D17. doi: 10.1093/nar/gkw1071.



Darling, J. A. and M. J. Blum (2007). “DNA-based methods for monitoring invasive species: a review and prospectus”. In: *Biological Invasions* 9.7, pp. 751–765. DOI: 10.1007/s10530-006-9079-4.

Dhariwal, A. et al. (2017). “MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data”. In: *Nucleic Acids Research* 45 (W1), W180–W188. DOI: 10.1093/nar/gkx295.

Edgar, R. C. (2010). “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19, pp. 2460–2461. DOI: 10.1093/bioinformatics/btq461. eprint: /oup/backfile/content_public/journal/bioinformatics/26/19/10.1093_bioinformatics_btq461/2/btq461.pdf.

EPPO (2017). *EPPO Global Database (available online).* URL: https://gd.eppo.int/ (visited on 09/22/2017).

Federhen, S. (2012). “The NCBI Taxonomy database”. In: *Nucleic Acids Research* 40 (D1), pp. D136–D143. DOI: 10.1093/nar/gkr1178.

Garibaldi, L. A. et al. (2013). “Wild Pollinators Enhance Fruit Set of Crops Regardless of Honey Bee Abundance”. In: *Science* 339.6127, pp. 1608–1611. DOI: 10.1126/science.1230200. eprint: http://science.sciencemag.org/content/339/6127/1608.full.pdf.

Hanley, N. et al. (2015). “Measuring the economic value of pollination services: Principles, evidence and knowledge gaps”. In: *Ecosystem Services* 14, pp. 124–132. DOI: 10.1016/j.ecoser.2014.09.013.

Henle, K. et al. (2014). “Scaling in Ecology and Biodiversity Conservation”. In: *Advanced Books* 1, e1169. DOI: 10.3897/ab.e1169.

Huse, S. M. et al. (2014). “VAMPS: a website for visualization and analysis of microbial population structures”. In: *BMC Bioinformatics* 15, p. 41. DOI: 10.1186/1471-2105-15-41.

Huson, D. H. et al. (2016). “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data”. In: *PLOS Computational Biology* 12.6, e1004957. DOI: 10.1371/journal.pcbi.1004957.

IUCN (2017). *IUCN Red List of Threatened Species. Version 2017-1.* URL: http://www.iucnredlist.org/ (visited on 09/22/2017).

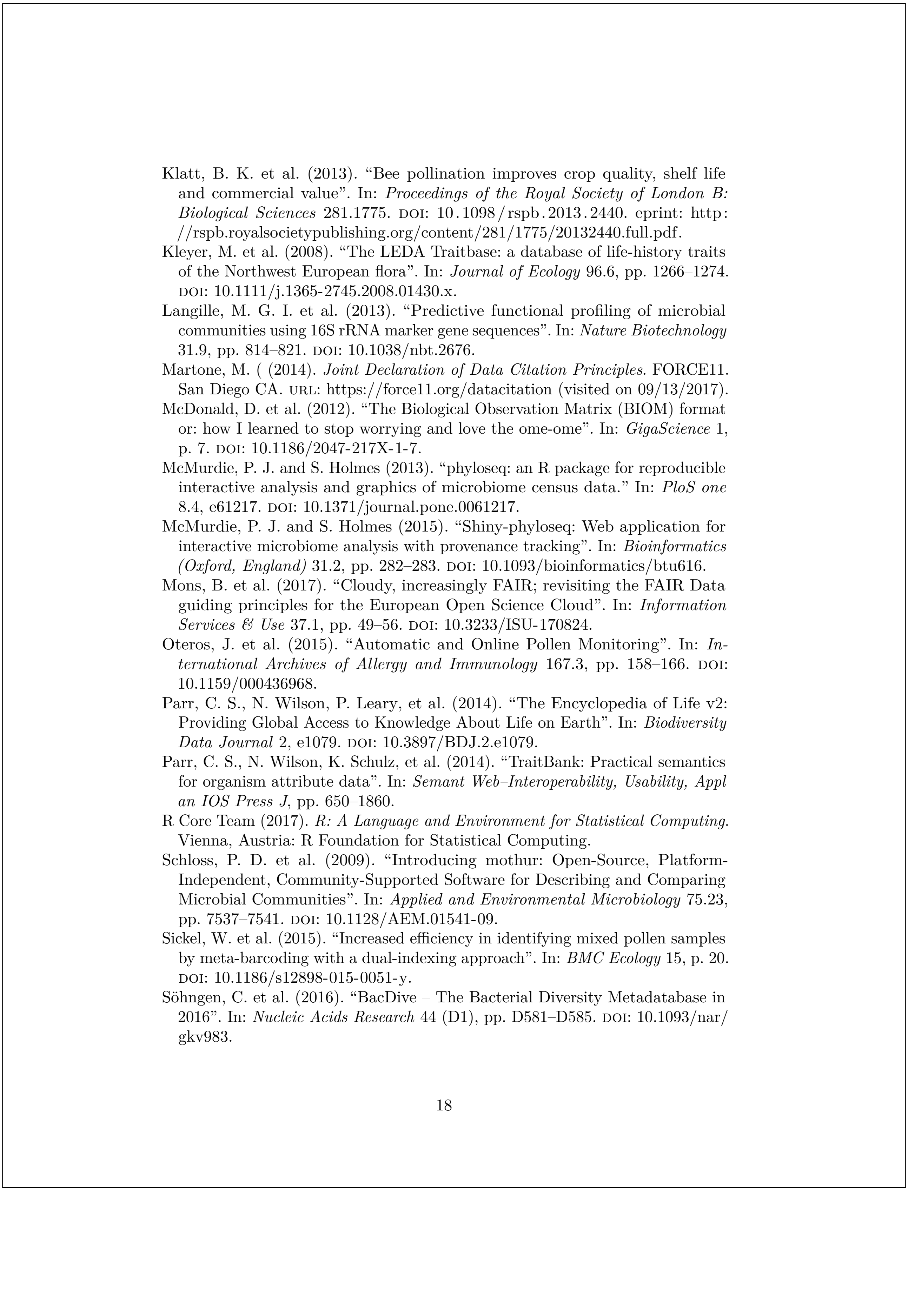
Junker, R. R., N. Blüthgen, and A. Keller (2015). “Functional and phylogenetic diversity of plant communities differently affect the structure of flower-visitor interactions and reveal convergences in floral traits”. In: *Evolutionary Ecology* 29.3, pp. 437–450. DOI: 10.1007/s10682-014-9747-2.

Kattge, J. et al. (2011). “TRY – a global database of plant traits”. In: *Global Change Biology* 17.9, pp. 2905–2935. DOI: 10.1111/j.1365-2486.2011.02451.x.

Katz, D. S. (2017). *FAIR is not fair enough.* Daniel S. Katz’s blog. URL: https://danielskatzblog.wordpress.com/2017/06/22/fair-is-not-fair-enough/ (visited on 09/13/2017).

Keller, A., N. Danner, et al. (2015). “Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples”. In: *Plant Biology* 17.2, pp. 558–566. DOI: 10.1111/plb.12251.

Keller, A., H. Horn, et al. (2014). “Computational integration of genomic traits into 16S rDNA microbiota sequencing studies”. In: *Gene* 549.1, pp. 186–191. DOI: 10.1016/j.gene.2014.07.066.



Klatt, B. K. et al. (2013). “Bee pollination improves crop quality, shelf life and commercial value”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 281.1775. DOI: 10.1098/rspb.2013.2440. eprint: http://rspb.royalsocietypublishing.org/content/281/1775/20132440.full.pdf.

Kleyer, M. et al. (2008). “The LEDA Traitbase: a database of life-history traits of the Northwest European flora”. In: *Journal of Ecology* 96.6, pp. 1266–1274. DOI: 10.1111/j.1365-2745.2008.01430.x.

Langille, M. G. I. et al. (2013). “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences”. In: *Nature Biotechnology* 31.9, pp. 814–821. DOI: 10.1038/nbt.2676.

Martone, M. ((2014). *Joint Declaration of Data Citation Principles.* FORCE11. San Diego CA. URL: https://force11.org/datacitation (visited on 09/13/2017).

McDonald, D. et al. (2012). “The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome”. In: *GigaScience* 1, p. 7. DOI: 10.1186/2047-217X-1-7.

McMurdie, P. J. and S. Holmes (2013). “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.” In: *PloS one* 8.4, e61217. DOI: 10.1371/journal.pone.0061217.

McMurdie, P. J. and S. Holmes (2015). “Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking”. In: *Bioinformatics (Oxford, England)* 31.2, pp. 282–283. DOI: 10.1093/bioinformatics/btu616.

Mons, B. et al. (2017). “Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud”. In: *Information Services & Use* 37.1, pp. 49–56. DOI: 10.3233/ISU-170824.

Oteros, J. et al. (2015). “Automatic and Online Pollen Monitoring”. In: *International Archives of Allergy and Immunology* 167.3, pp. 158–166. DOI: 10.1159/000436968.

Parr, C. S., N. Wilson, P. Leary, et al. (2014). “The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth”. In: *Biodiversity Data Journal* 2, e1079. DOI: 10.3897/BDJ.2.e1079.

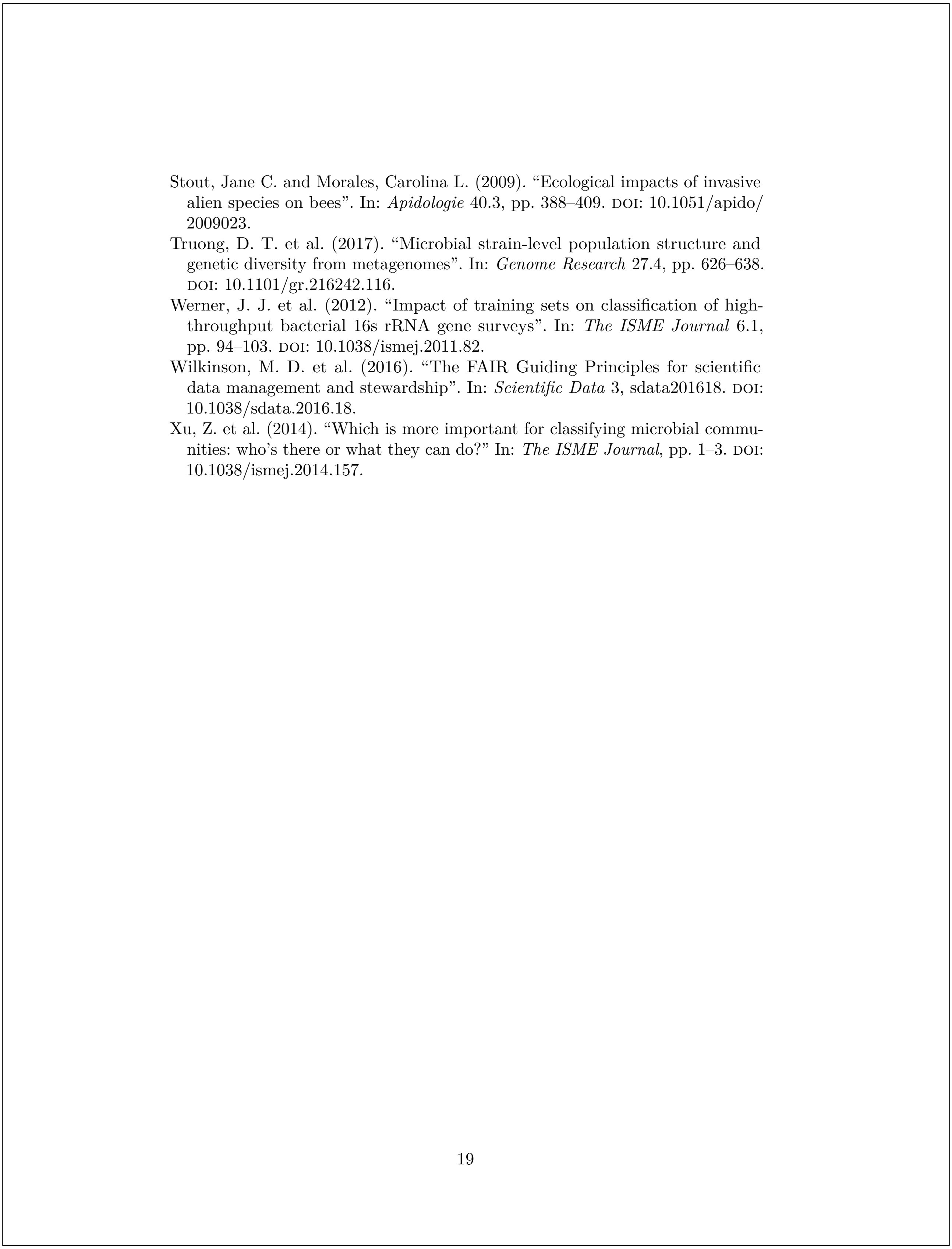
Parr, C. S., N. Wilson, K. Schulz, et al. (2014). “TraitBank: Practical semantics for organism attribute data”. In: *Semant Web–Interoperability, Usability, Appl an IOS Press J*, pp. 650–1860.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schloss, P. D. et al. (2009). “Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities”. In: *Applied and Environmental Microbiology* 75.23, pp. 7537–7541. DOI: 10.1128/AEM.01541-09.

Sickel, W. et al. (2015). “Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach”. In: *BMC Ecology* 15, p. 20. DOI: 10.1186/s12898-015-0051-y.

Söhngen, C. et al. (2016). “BacDive – The Bacterial Diversity Metadatabase in 2016”. In: *Nucleic Acids Research* 44 (D1), pp. D581–D585. DOI: 10.1093/nar/gkv983.



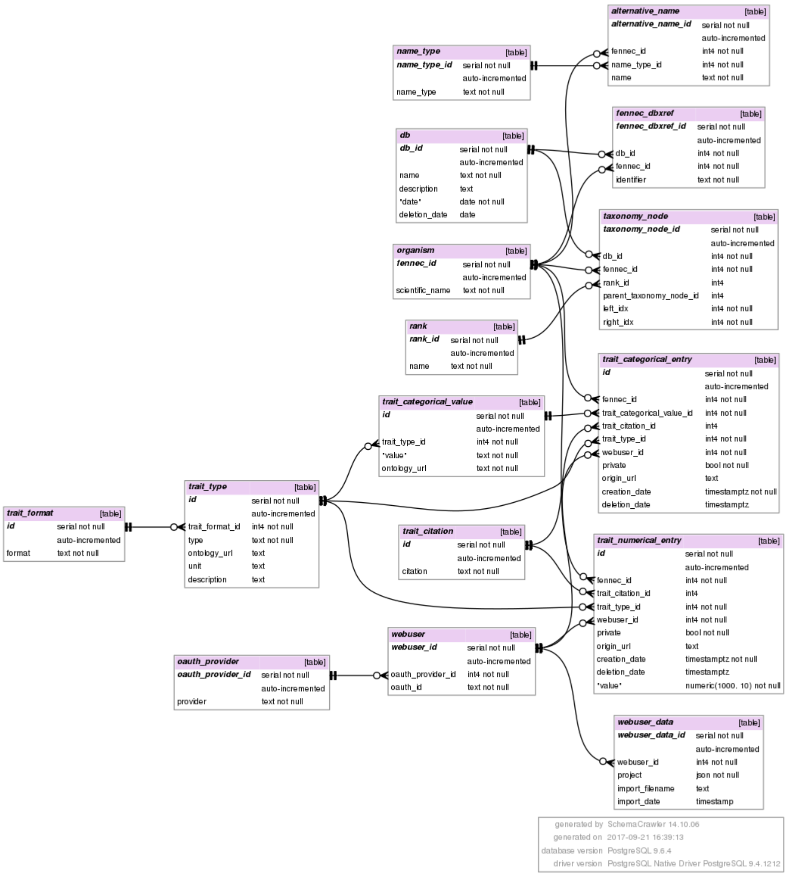
Stout, Jane C. and Morales, Carolina L. (2009). “Ecological impacts of invasive alien species on bees”. In: *Apidologie* 40.3, pp. 388–409. DOI: 10.1051/apido/2009023.

Truong, D. T. et al. (2017). “Microbial strain-level population structure and genetic diversity from metagenomes”. In: *Genome Research* 27.4, pp. 626–638. DOI: 10.1101/gr.216242.116.

Werner, J. J. et al. (2012). “Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys”. In: *The ISME Journal* 6.1, pp. 94–103. DOI: 10.1038/ismej.2011.82.

Wilkinson, M. D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3, sdata201618. DOI: 10.1038/sdata.2016.18.

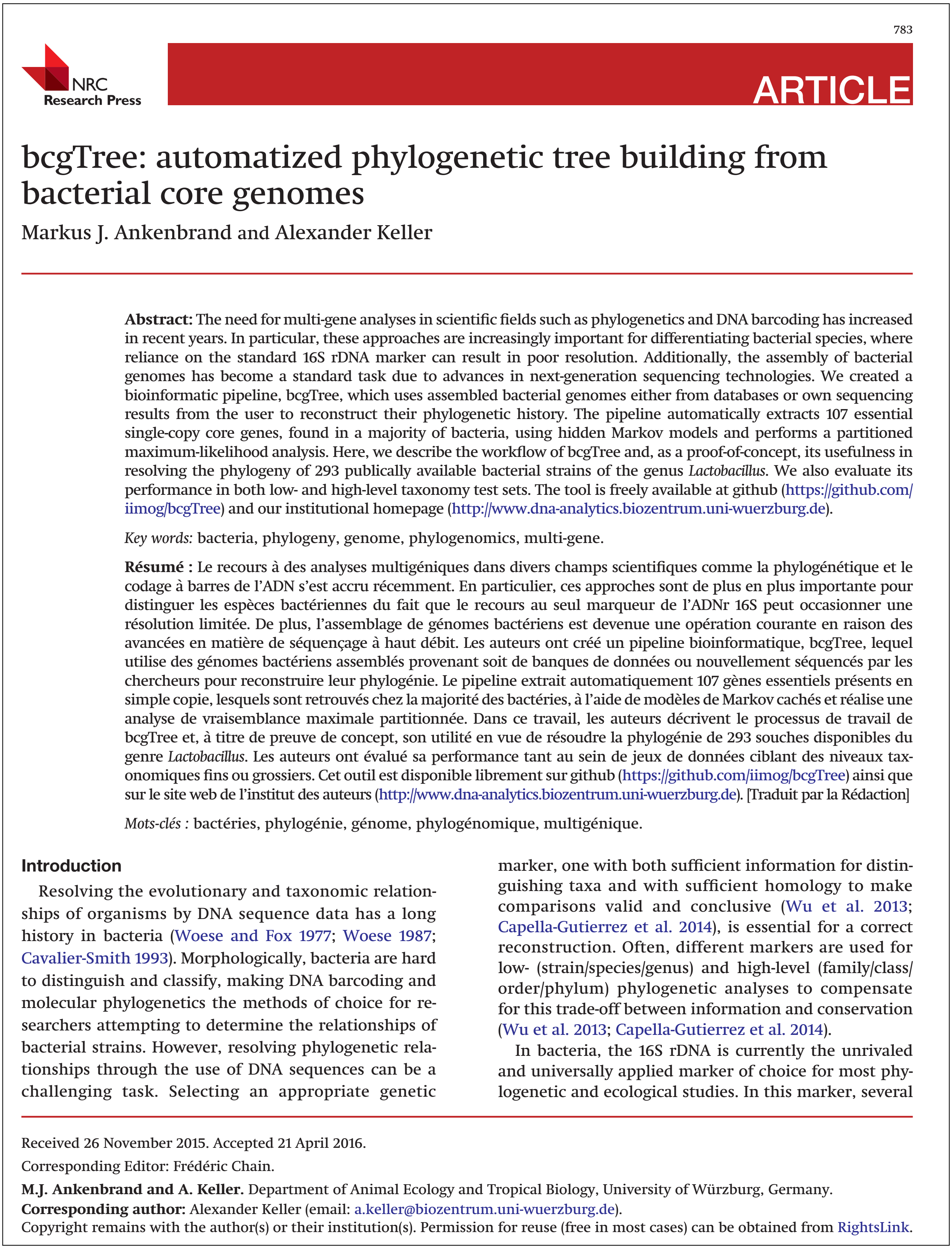
Xu, Z. et al. (2014). “Which is more important for classifying microbial communities: who’s there or what they can do?” In: *The ISME Journal*, pp. 1–3. DOI: 10.1038/ismej.2014.157.


name_type [table]
name_type_id serial not null auto-incremented
name_type text not null
alternative_name [table]
alternative_name_id serial not null auto-incremented
fennec_id int4 not null
name_type_id int4 not null
name text not null
db [table]
db_id serial not null auto-incremented
name text not null
description text
"date" date not null
deletion_date date
fennec_dbxref [table]
fennec_dbxref_id serial not null auto-incremented
db_id int4 not null
fennec_id int4 not null
identifier text not null
organism [table]
fennec_id serial not null auto-incremented
scientific_name text not null
taxonomy_node [table]
taxonomy_node_id serial not null auto-incremented
db_id int4 not null
fennec_id int4 not null
rank_id int4
parent_taxonomy_node_id int4
left_idx int4 not null
right_idx int4 not null
rank [table]
rank_id serial not null auto-incremented
name text not null
trait_categorical_value [table]
id serial not null auto-incremented
trait_type_id int4 not null
"value" text not null
ontology_url text not null
trait_categorical_entry [table]
id serial not null auto-incremented
fennec_id int4 not null
trait_categorical_value_id int4 not null
trait_citation_id int4
trait_type_id int4 not null
webuser_id int4 not null
private bool not null
origin_url text
creation_date timestamptz not null
deletion_date timestamptz
trait_format [table]
id serial not null auto-incremented
format text not null
trait_type [table]
id serial not null auto-incremented
trait_format_id int4 not null
type text not null
ontology_url text
unit text
description text
trait_citation [table]
id serial not null auto-incremented
citation text not null
trait_numerical_entry [table]
id serial not null auto-incremented
fennec_id int4 not null
trait_citation_id int4
trait_type_id int4 not null
webuser_id int4 not null
private bool not null
origin_url text
creation_date timestamptz not null
deletion_date timestamptz
"value" numeric(1000. 10) not null
oauth_provider [table]
oauth_provider_id serial not null auto-incremented
provider text not null
webuser [table]
webuser_id serial not null auto-incremented
oauth_provider_id int4 not null
oauth_id text not null
webuser_data [table]
webuser_data_id serial not null auto-incremented
webuser_id int4 not null
project json not null
import_filename text
import_date timestamp
generated by SchemaCrawler 14.10.06
generated on 2017-09-21 16:39:13
database version PostgreSQL 9.6.4
driver version PostgreSQL Native Driver PostgreSQL 9.4.1212

Supplemental Figure 1: Schema of the Fennec database, created with SchemaCrawler 14.10.06 (https://github.com/sualeh/SchemaCrawler).

5.8 BCGTREE: AUTOMATIZED PHYLOGENETIC TREE BUILDING FROM BACTERIAL CORE GENOMES

– published in *Genome* –



783

NRC Research Press

ARTICLE

bcgTree: automatized phylogenetic tree building from bacterial core genomes

Markus J. Ankenbrand and Alexander Keller

Abstract: The need for multi-gene analyses in scientific fields such as phylogenetics and DNA barcoding has increased in recent years. In particular, these approaches are increasingly important for differentiating bacterial species, where reliance on the standard 16S rDNA marker can result in poor resolution. Additionally, the assembly of bacterial genomes has become a standard task due to advances in next-generation sequencing technologies. We created a bioinformatic pipeline, bcgTree, which uses assembled bacterial genomes either from databases or own sequencing results from the user to reconstruct their phylogenetic history. The pipeline automatically extracts 107 essential single-copy core genes, found in a majority of bacteria, using hidden Markov models and performs a partitioned maximum-likelihood analysis. Here, we describe the workflow of bcgTree and, as a proof-of-concept, its usefulness in resolving the phylogeny of 293 publically available bacterial strains of the genus *Lactobacillus*. We also evaluate its performance in both low- and high-level taxonomy test sets. The tool is freely available at github (https://github.com/iimog/bcgTree) and our institutional homepage (http://www.dna-analytics.biozentrum.uni-wuerzburg.de).

Key words: bacteria, phylogeny, genome, phylogenomics, multi-gene.

Résumé : Le recours à des analyses multigéniques dans divers champs scientifiques comme la phylogénétique et le codage à barres de l'ADN s'est accru récemment. En particulier, ces approches sont de plus en plus importante pour distinguer les espèces bactériennes du fait que le recours au seul marqueur de l'ADNr 16S peut occasionner une résolution limitée. De plus, l'assemblage de génomes bactériens est devenue une opération courante en raison des avancées en matière de séquençage à haut débit. Les auteurs ont créé un pipeline bioinformatique, bcgTree, lequel utilise des génomes bactériens assemblés provenant soit de banques de données ou nouvellement séquencés par les chercheurs pour reconstruire leur phylogénie. Le pipeline extrait automatiquement 107 gènes essentiels présents en simple copie, lesquels sont retrouvés chez la majorité des bactéries, à l'aide de modèles de Markov cachés et réalise une analyse de vraisemblance maximale partitionnée. Dans ce travail, les auteurs décrivent le processus de travail de bcgTree et, à titre de preuve de concept, son utilité en vue de résoudre la phylogénie de 293 souches disponibles du genre *Lactobacillus*. Les auteurs ont évalué sa performance tant au sein de jeux de données ciblant des niveaux taxonomiques fins ou grossiers. Cet outil est disponible librement sur github (https://github.com/iimog/bcgTree) ainsi que sur le site web de l'institut des auteurs (http://www.dna-analytics.biozentrum.uni-wuerzburg.de). [Traduit par la Rédaction]

Mots-clés : bactéries, phylogénie, génome, phylogénomique, multigénique.

Introduction

Resolving the evolutionary and taxonomic relationships of organisms by DNA sequence data has a long history in bacteria (Woese and Fox 1977; Woese 1987; Cavalier-Smith 1993). Morphologically, bacteria are hard to distinguish and classify, making DNA barcoding and molecular phylogenetics the methods of choice for researchers attempting to determine the relationships of bacterial strains. However, resolving phylogenetic relationships through the use of DNA sequences can be a challenging task. Selecting an appropriate genetic marker, one with both sufficient information for distinguishing taxa and with sufficient homology to make comparisons valid and conclusive (Wu et al. 2013; Capella-Gutierrez et al. 2014), is essential for a correct reconstruction. Often, different markers are used for low- (strain/species/genus) and high-level (family/class/order/phylum) phylogenetic analyses to compensate for this trade-off between information and conservation (Wu et al. 2013; Capella-Gutierrez et al. 2014).

In bacteria, the 16S rDNA is currently the unrivaled and universally applied marker of choice for most phylogenetic and ecological studies. In this marker, several

Received 26 November 2015. Accepted 21 April 2016.

Corresponding Editor: Frédéric Chain.

M.J. Ankenbrand and A. Keller. Department of Animal Ecology and Tropical Biology, University of Würzburg, Germany.
Corresponding author: Alexander Keller (email: a.keller@biozentrum.uni-wuerzburg.de).

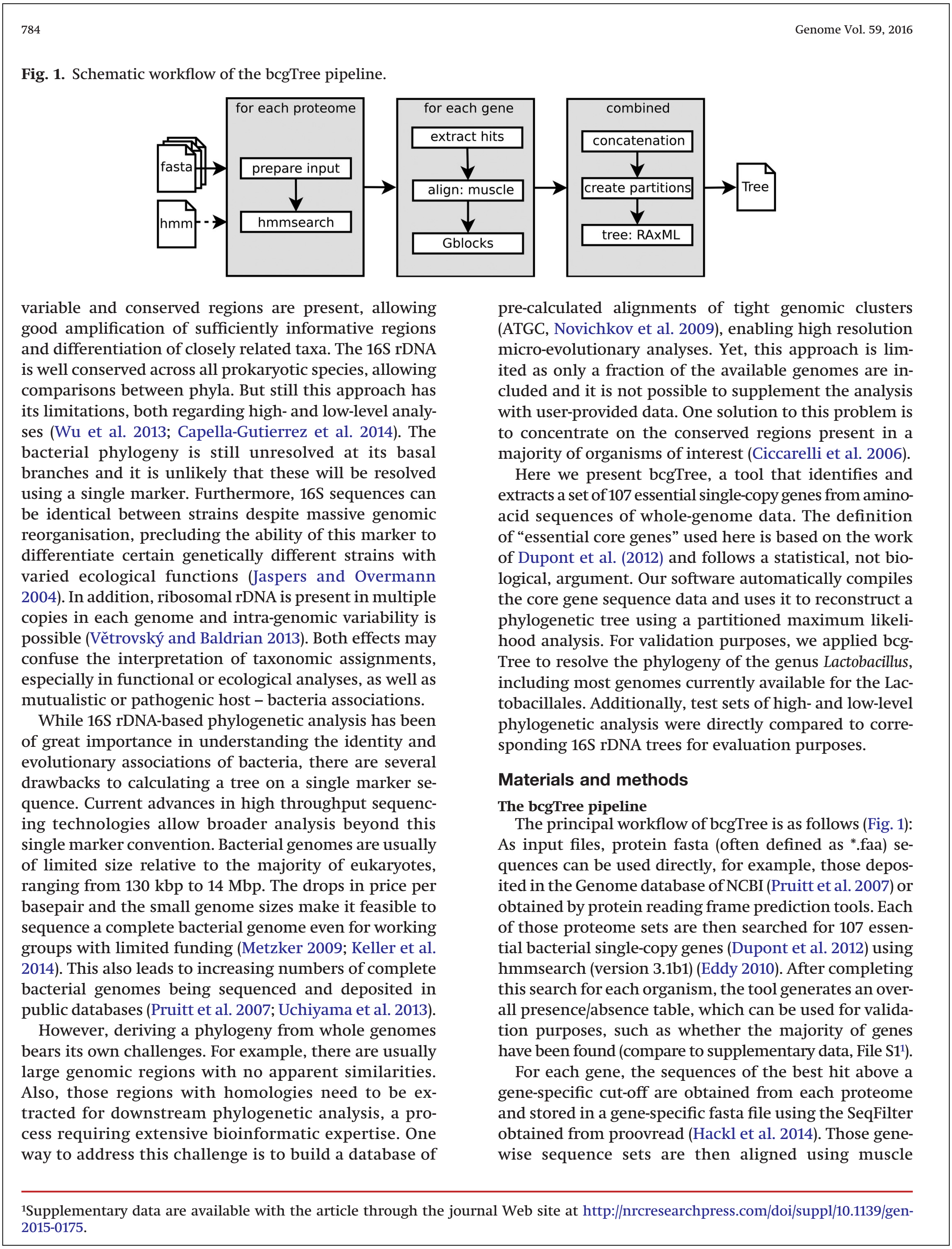


Fig. 1. Schematic workflow of the bcgTree pipeline.

for each proteome
fasta
hmm
prepare input
hmmsearch
for each gene
extract hits
align: muscle
Gblocks
combined
concatenation
create partitions
tree: RAxML
Tree

variable and conserved regions are present, allowing good amplification of sufficiently informative regions and differentiation of closely related taxa. The 16S rDNA is well conserved across all prokaryotic species, allowing comparisons between phyla. But still this approach has its limitations, both regarding high- and low-level analyses (Wu et al. 2013; Capella-Gutierrez et al. 2014). The bacterial phylogeny is still unresolved at its basal branches and it is unlikely that these will be resolved using a single marker. Furthermore, 16S sequences can be identical between strains despite massive genomic reorganisation, precluding the ability of this marker to differentiate certain genetically different strains with varied ecological functions (Jaspers and Overmann 2004). In addition, ribosomal rDNA is present in multiple copies in each genome and intra-genomic variability is possible (Větrovský and Baldrian 2013). Both effects may confuse the interpretation of taxonomic assignments, especially in functional or ecological analyses, as well as mutualistic or pathogenic host – bacteria associations.

While 16S rDNA-based phylogenetic analysis has been of great importance in understanding the identity and evolutionary associations of bacteria, there are several drawbacks to calculating a tree on a single marker sequence. Current advances in high throughput sequencing technologies allow broader analysis beyond this single marker convention. Bacterial genomes are usually of limited size relative to the majority of eukaryotes, ranging from 130 kbp to 14 Mbp. The drops in price per basepair and the small genome sizes make it feasible to sequence a complete bacterial genome even for working groups with limited funding (Metzker 2009; Keller et al. 2014). This also leads to increasing numbers of complete bacterial genomes being sequenced and deposited in public databases (Pruitt et al. 2007; Uchiyama et al. 2013).

However, deriving a phylogeny from whole genomes bears its own challenges. For example, there are usually large genomic regions with no apparent similarities. Also, those regions with homologies need to be extracted for downstream phylogenetic analysis, a process requiring extensive bioinformatic expertise. One way to address this challenge is to build a database of pre-calculated alignments of tight genomic clusters (ATGC, Novichkov et al. 2009), enabling high resolution micro-evolutionary analyses. Yet, this approach is limited as only a fraction of the available genomes are included and it is not possible to supplement the analysis with user-provided data. One solution to this problem is to concentrate on the conserved regions present in a majority of organisms of interest (Ciccarelli et al. 2006).

Here we present bcgTree, a tool that identifies and extracts a set of 107 essential single-copy genes from amino-acid sequences of whole-genome data. The definition of "essential core genes" used here is based on the work of Dupont et al. (2012) and follows a statistical, not biological, argument. Our software automatically compiles the core gene sequence data and uses it to reconstruct a phylogenetic tree using a partitioned maximum likelihood analysis. For validation purposes, we applied bcgTree to resolve the phylogeny of the genus *Lactobacillus*, including most genomes currently available for the Lactobacillales. Additionally, test sets of high- and low-level phylogenetic analysis were directly compared to corresponding 16S rDNA trees for evaluation purposes.

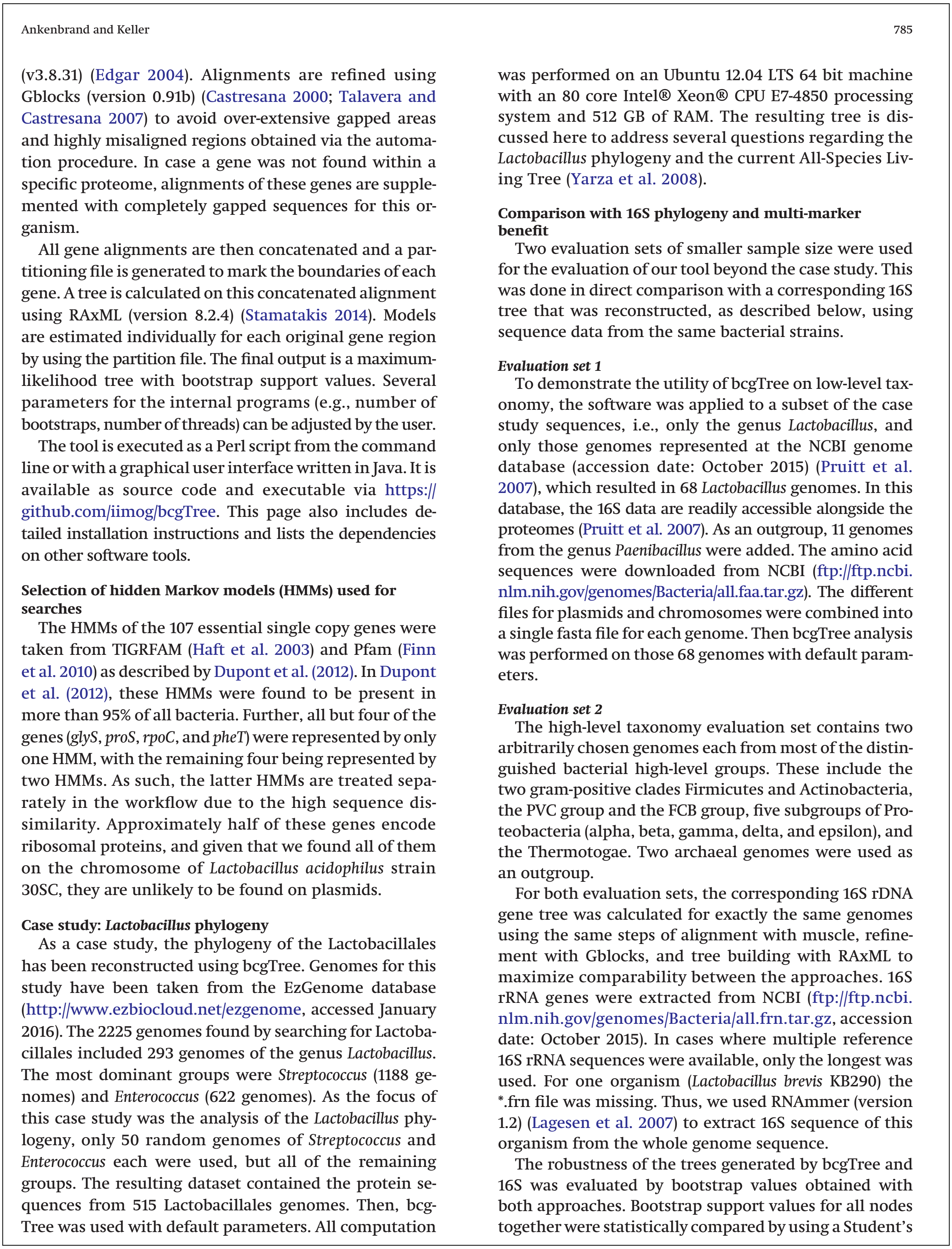
Materials and methods

The bcgTree pipeline

The principal workflow of bcgTree is as follows (Fig. 1): As input files, protein fasta (often defined as *.faa) sequences can be used directly, for example, those deposited in the Genome database of NCBI (Pruitt et al. 2007) or obtained by protein reading frame prediction tools. Each of those proteome sets are then searched for 107 essential bacterial single-copy genes (Dupont et al. 2012) using hmmsearch (version 3.1b1) (Eddy 2010). After completing this search for each organism, the tool generates an overall presence/absence table, which can be used for validation purposes, such as whether the majority of genes have been found (compare to supplementary data, File S1[1]).

For each gene, the sequences of the best hit above a gene-specific cut-off are obtained from each proteome and stored in a gene-specific fasta file using the SeqFilter obtained from proovread (Hackl et al. 2014). Those gene-wise sequence sets are then aligned using muscle

[1] Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2015-0175.



(v3.8.31) (Edgar 2004). Alignments are refined using Gblocks (version 0.91b) (Castresana 2000; Talavera and Castresana 2007) to avoid over-extensive gapped areas and highly misaligned regions obtained via the automation procedure. In case a gene was not found within a specific proteome, alignments of these genes are supplemented with completely gapped sequences for this organism.

All gene alignments are then concatenated and a partitioning file is generated to mark the boundaries of each gene. A tree is calculated on this concatenated alignment using RAxML (version 8.2.4) (Stamatakis 2014). Models are estimated individually for each original gene region by using the partition file. The final output is a maximum-likelihood tree with bootstrap support values. Several parameters for the internal programs (e.g., number of bootstraps, number of threads) can be adjusted by the user.

The tool is executed as a Perl script from the command line or with a graphical user interface written in Java. It is available as source code and executable via https://github.com/iimog/bcgTree. This page also includes detailed installation instructions and lists the dependencies on other software tools.

Selection of hidden Markov models (HMMs) used for searches

The HMMs of the 107 essential single copy genes were taken from TIGRFAM (Haft et al. 2003) and Pfam (Finn et al. 2010) as described by Dupont et al. (2012). In Dupont et al. (2012), these HMMs were found to be present in more than 95% of all bacteria. Further, all but four of the genes (*glyS*, *proS*, *rpoC*, and *pheT*) were represented by only one HMM, with the remaining four being represented by two HMMs. As such, the latter HMMs are treated separately in the workflow due to the high sequence dissimilarity. Approximately half of these genes encode ribosomal proteins, and given that we found all of them on the chromosome of *Lactobacillus acidophilus* strain 30SC, they are unlikely to be found on plasmids.

Case study: *Lactobacillus* phylogeny

As a case study, the phylogeny of the Lactobacillales has been reconstructed using bcgTree. Genomes for this study have been taken from the EzGenome database (http://www.ezbiocloud.net/ezgenome, accessed January 2016). The 2225 genomes found by searching for Lactobacillales included 293 genomes of the genus *Lactobacillus*. The most dominant groups were *Streptococcus* (1188 genomes) and *Enterococcus* (622 genomes). As the focus of this case study was the analysis of the *Lactobacillus* phylogeny, only 50 random genomes of *Streptococcus* and *Enterococcus* each were used, but all of the remaining groups. The resulting dataset contained the protein sequences from 515 Lactobacillales genomes. Then, bcgTree was used with default parameters. All computation was performed on an Ubuntu 12.04 LTS 64 bit machine with an 80 core Intel® Xeon® CPU E7-4850 processing system and 512 GB of RAM. The resulting tree is discussed here to address several questions regarding the *Lactobacillus* phylogeny and the current All-Species Living Tree (Yarza et al. 2008).

Comparison with 16S phylogeny and multi-marker benefit

Two evaluation sets of smaller sample size were used for the evaluation of our tool beyond the case study. This was done in direct comparison with a corresponding 16S tree that was reconstructed, as described below, using sequence data from the same bacterial strains.

Evaluation set 1

To demonstrate the utility of bcgTree on low-level taxonomy, the software was applied to a subset of the case study sequences, i.e., only the genus *Lactobacillus*, and only those genomes represented at the NCBI genome database (accession date: October 2015) (Pruitt et al. 2007), which resulted in 68 *Lactobacillus* genomes. In this database, the 16S data are readily accessible alongside the proteomes (Pruitt et al. 2007). As an outgroup, 11 genomes from the genus *Paenibacillus* were added. The amino acid sequences were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.faa.tar.gz). The different files for plasmids and chromosomes were combined into a single fasta file for each genome. Then bcgTree analysis was performed on those 68 genomes with default parameters.

Evaluation set 2

The high-level taxonomy evaluation set contains two arbitrarily chosen genomes each from most of the distinguished bacterial high-level groups. These include the two gram-positive clades Firmicutes and Actinobacteria, the PVC group and the FCB group, five subgroups of Proteobacteria (alpha, beta, gamma, delta, and epsilon), and the Thermotogae. Two archaeal genomes were used as an outgroup.

For both evaluation sets, the corresponding 16S rDNA gene tree was calculated for exactly the same genomes using the same steps of alignment with muscle, refinement with Gblocks, and tree building with RAxML to maximize comparability between the approaches. 16S rRNA genes were extracted from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.frn.tar.gz, accession date: October 2015). In cases where multiple reference 16S rRNA sequences were available, only the longest was used. For one organism (*Lactobacillus brevis* KB290) the *.frn file was missing. Thus, we used RNAmmer (version 1.2) (Lagesen et al. 2007) to extract 16S sequence of this organism from the whole genome sequence.

The robustness of the trees generated by bcgTree and 16S was evaluated by bootstrap values obtained with both approaches. Bootstrap support values for all nodes together were statistically compared by using a Student's

Fig. 2. Case study tree calculated with bcgTree containing 515 Lactobacillales genomes. Numbers at nodes designate bootstrap support values resulting from 100 bootstrap replicates. Outgroup is *Aerococcus*. Monophyletic genera and species have been collapsed as <G> and <S>, respectively, (represented as a triangle considering intra-group variation as distance, with number of included genomes in curly brackets).

t test in R (R Development Core Team 2010). Tree topologies for both evaluation sets were compared between bcgTree and 16S using the R package dendextend (Galili 2015) for tanglegrams. Differences in topologies were highlighted using the same package with dashed lines.

The influence of the number of genes used for the tree-building process on the accuracy of the resulting tree was also assessed. For this, we used the final alignment files obtained through bcgTree for both evaluation sets with all 109 partitions (two of the 107 genes have two partitions each). These were randomly subsampled using RAxML and their corresponding partitions in the alignment excluded. For both evaluation sets, we used this approach to create concatenated alignment files with 1 to 108 random HMMs with 10 replicates per number. RAxML-derived trees were constructed using these alignments without bootstrapping and with the same parameters used in the default bcgTree analysis. The quartet distances between the resulting trees and the full gene set tree was calculated using qdist (version 2.0) (Mailund and Pedersen 2004). The quartet distance is a measure of the topological distance between two phylogenetic trees (Mailund and Pedersen 2004; Keller et al. 2010). For visualization purposes, the number of genes was rounded to increments of five. The quartet distance of the 16S rDNA tree was also calculated for comparison.

Computational performance

The computational performance of bcgTree on sets with different numbers of complete genomes was assessed on a standard dual-core desktop computer (Intel® Core$^{\text{TM}}$2 Duo CPU E8500, 4 GB RAM, Ubuntu 14.04.3 LTS 64 bit). Since bcgTree is designed to work on complete genomes and has a fixed set of 107 essential genes, the only variable is the number of genomes. Genome size variation is not expected to change the overall runtime substantially and was not evaluated.

The pipeline was executed on a variable number of genomes: for 5–15 genomes, each step-size of one was repeated with five replicates, for 5–50 genomes only one replicate was done with a step-size of five. For all steps, random proteomes were selected from all data downloaded from the genome database of NCBI. The total runtime of bcgTree was further separated into time before and after start of RAxML, to estimate the proportion of preparation and tree calculation of the total runtime. Linearity of the runtime increase with number of genomes was tested using a linear model for complete and pre-RAxML runtime.

Results and discussion

Case study: *Lactobacillus* phylogeny

The case-study tree automatically generated by bcgTree largely supports the monophyly of most genera within the Lactobacillales (Fig. 2). However, *Pediococcus* (9 genomes) and the family Leuconostocaceae (51 genomes that form a monophylum and include the genera *Weissella*, *Oenococcus*, *Fructobacillus*, *Leuconostoc*) get inserted into the *Lactobacillus* genus, thus violating the monophyly of the latter. This is a known phenomenon that can also be observed in the All-Species Living Tree (Yarza et al. 2008). Within the genus *Lactobacillus*, the tree is well resolved and consistent with previously published results on the genus (Kant et al. 2011). Most species are well resolved into monophyletic clusters with high support values, thus providing better assignments than 16S only based analyses (Yarza et al. 2008), yet the All-Species Living Tree contains only a single representative of most *Lactobacillus* species.

The bcgTree phylogeny also supports the three major groups of *Lactobacillus* species as listed in *Bergey's Manual of Systematic Bacteriology* (Vol. 3) (De Vos et al. 2011) and Kant et al. (2011) (*L. delbrueckii* group or NCFM, *L. reuteri* group or WCFS1, *L. salivarius* group or WCFS2, and *L. rhamnosus/casei* group or GG) as monophyla with some new species added. The WCFS group reported by Kant et al. (2011) is split in the bcgTree tree into two groups separated by the Leuconostocaceae, which were not considered in that study. Furthermore, there are *Lactobacillus* strains that do not belong to any of these groups. Within the *L. rhamnosus/casei* group, *L. casei* (22 genomes) and *L. paracasei* (37 genomes), the leaves of the tree were highly intermixed, suggesting that it might not be appropriate to assign these to different species.

Comparison to 16S topology

For the low-taxonomy evaluation set of *Lactobacillus*, bootstrap support values were consistently higher with bcgTree than 16S data only (File S2[1], t = 2.25, df = 27, p < 0.05*; Fig. 3). The presence/absence results of the bcgTree procedure with *Lactobacillus* show two genomes that lack a high proportion of the included genes (File S1[1]). These were *Lactobacillus fermentum* CECT 5716 and *Lactobacillus salviarius* CECT 5713. Both genomes are of typical size, suggesting that they have been well assembled. They showed, however, a reduction by approximately one half of predicted open reading frames, indicating incomplete annotation. Thus the proteome files were smaller than those of closely related taxa. Still the tree reconstruction with bcgTree assigned them adequately at their expected positions in the trees, and it

Ankenbrand and Keller 787

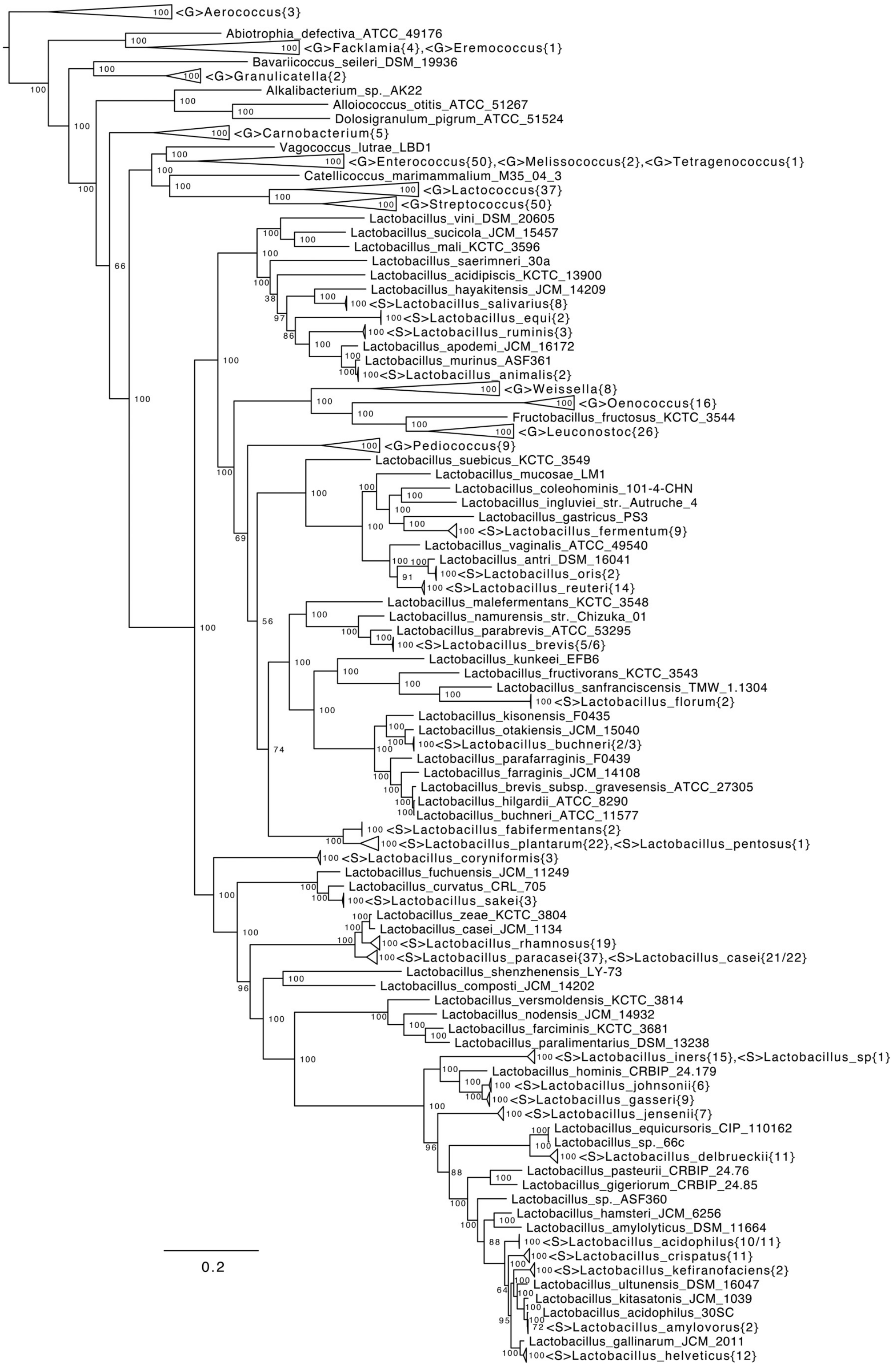
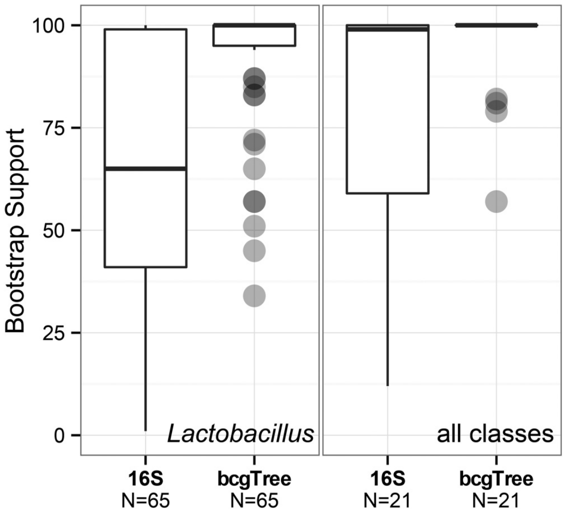

<G>Aerococcus{3}
Abiotrophia_defectiva_ATCC_49176
<G>Facklamia{4},<G>Eremococcus{1}
Bavariicoccus_seileri_DSM_19936
<G>Granulicatella{2}
Alkalibacterium_sp._AK22
Alloiococcus_otitis_ATCC_51267
Dolosigranulum_pigrum_ATCC_51524
<G>Carnobacterium{5}
Vagococcus_lutrae_LBD1
<G>Enterococcus{50},<G>Melissococcus{2},<G>Tetragenococcus{1}
Catellicoccus_marimammalium_M35_04_3
<G>Lactococcus{37}
<G>Streptococcus{50}
Lactobacillus_vini_DSM_20605
Lactobacillus_sucicola_JCM_15457
Lactobacillus_mali_KCTC_3596
Lactobacillus_saerimneri_30a
Lactobacillus_acidipiscis_KCTC_13900
Lactobacillus_hayakitensis_JCM_14209
<S>Lactobacillus_salivarius{8}
<S>Lactobacillus_equi{2}
<S>Lactobacillus_ruminis{3}
Lactobacillus_apodemi_JCM_16172
Lactobacillus_murinus_ASF361
<S>Lactobacillus_animalis{2}
<G>Weissella{8}
<G>Oenococcus{16}
Fructobacillus_fructosus_KCTC_3544
<G>Leuconostoc{26}
<G>Pediococcus{9}
Lactobacillus_suebicus_KCTC_3549
Lactobacillus_mucosae_LM1
Lactobacillus_coleohominis_101-4-CHN
Lactobacillus_ingluviei_str._Autruche_4
Lactobacillus_gastricus_PS3
<S>Lactobacillus_fermentum{9}
Lactobacillus_vaginalis_ATCC_49540
Lactobacillus_antri_DSM_16041
<S>Lactobacillus_oris{2}
<S>Lactobacillus_reuteri{14}
Lactobacillus_malefermentans_KCTC_3548
Lactobacillus_namurensis_str._Chizuka_01
Lactobacillus_parabrevis_ATCC_53295
<S>Lactobacillus_brevis{5/6}
Lactobacillus_kunkeei_EFB6
Lactobacillus_fructivorans_KCTC_3543
Lactobacillus_sanfranciscensis_TMW_1.1304
<S>Lactobacillus_florum{2}
Lactobacillus_kisonensis_F0435
Lactobacillus_otakiensis_JCM_15040
<S>Lactobacillus_buchneri{2/3}
Lactobacillus_parafarraginis_F0439
Lactobacillus_farraginis_JCM_14108
Lactobacillus_brevis_subsp._gravesensis_ATCC_27305
Lactobacillus_hilgardii_ATCC_8290
Lactobacillus_buchneri_ATCC_11577
<S>Lactobacillus_fabifermentans{2}
<S>Lactobacillus_plantarum{22},<S>Lactobacillus_pentosus{1}
<S>Lactobacillus_coryniformis{3}
Lactobacillus_fuchuensis_JCM_11249
Lactobacillus_curvatus_CRL_705
<S>Lactobacillus_sakei{3}
Lactobacillus_zeae_KCTC_3804
Lactobacillus_casei_JCM_1134
<S>Lactobacillus_rhamnosus{19}
<S>Lactobacillus_paracasei{37},<S>Lactobacillus_casei{21/22}
Lactobacillus_shenzhenensis_LY-73
Lactobacillus_composti_JCM_14202
Lactobacillus_versmoldensis_KCTC_3814
Lactobacillus_nodensis_JCM_14932
Lactobacillus_farciminis_KCTC_3681
Lactobacillus_paralimentarius_DSM_13238
<S>Lactobacillus_iners{15},<S>Lactobacillus_sp{1}
Lactobacillus_hominis_CRBIP_24.179
<S>Lactobacillus_johnsonii{6}
<S>Lactobacillus_gasseri{9}
<S>Lactobacillus_jensenii{7}
Lactobacillus_equicursoris_CIP_110162
Lactobacillus_sp._66c
<S>Lactobacillus_delbrueckii{11}
Lactobacillus_pasteurii_CRBIP_24.76
Lactobacillus_gigeriorum_CRBIP_24.85
Lactobacillus_sp._ASF360
Lactobacillus_hamsteri_JCM_6256
Lactobacillus_amylolyticus_DSM_11664
<S>Lactobacillus_acidophilus{10/11}
<S>Lactobacillus_crispatus{11}
<S>Lactobacillus_kefiranofaciens{2}
Lactobacillus_ultunensis_DSM_16047
Lactobacillus_kitasatonis_JCM_1039
Lactobacillus_acidophilus_30SC
<S>Lactobacillus_amylovorus{2}
Lactobacillus_gallinarum_JCM_2011
<S>Lactobacillus_helveticus{12}
0.2

Fig. 3. Boostrap support values for bcgTree and 16S rDNA only for direct comparison. Both evaluation sets, i.e., high- and low-level taxonomy are displayed.



was robust with high bootstrap values. This indicates that the tool is resilient to missing gene predictions in the proteome files. Also, some genes are represented by two HMMs and in almost all cases a hit was found for only one HMM per gene.

For the second evaluation set on high-level taxonomy, the bcgTree tree outperformed the 16S rDNA tree in terms of robustness with bootstrap support values (File S3[1], t = 6.31, df = 93, $p < 0.001^{***}$; Fig. 3). The general topology of the bcgTree tree was in concordance with the currently prevailing opinion of bacterial phylogeny. The bcgTree tree results provided support for monophyly of the two gram-positive groups Firmicutes and Actinobacteria. These groups were not resolved using 16S sequences alone. Also the Spirochaetes, PVC and FCB cluster together only with bcgTree, which is the current consensus opinion of the relatedness of these clades according to Bergey's Taxonomic Outlines (Ludwig et al. 2010). The remaining clades were resolved consistently with both methods, although with slightly different arrangements between groups. Please consider, however, that this study does not intend to resolve the molecular phylogeny of bacterial high-level groups and that taxa were arbitrarily chosen to validate the utility of the bcgTree approach for higher-level taxonomy.

Multi-marker benefits

For both evaluation sets, we subsampled the numbers of genes randomly and compared the resulting trees with the tree calculated on the complete gene set to infer influences of gene number on the accuracy of the method. In both evaluation sets, the quartet distance of trees calculated on subsets of the genes strongly decreases with an increase in the number of genes included (see Fig. 4). This shows that including more genes has great impact and benefits the accuracy of obtaining a tree similar to the full gene set. In both sets it can be seen that including five genes already leads to a great improvement in comparison to single gene analyses and this benefit appears to continually increase with more genes. The quartet distance of the 16S rDNA *Lactobacillus* tree is higher than most of the bcgTree-derived trees calculated with even small subsets of the essential genes. In contrast for the high-level taxonomy example, the 16S rDNA tree has a lower quartet distance than most of the small subset trees but a higher quartet distance than the bcgTree trees constructed using the full 107 genes and large subsets of the genes. This observation highlights the suitability of 16S rDNA as a marker for high-level taxonomy while demonstrating that the single-marker 16S rDNA analyses can be improved upon through multi-marker approaches, such as bcgTree.

For our case study the mean number and standard deviation of genes identified and used per genome was 104.0 ± 6.4. For the low-level as well as the high-level evaluation sets this was 104.9 ± 9.0 and 99.4 ± 21.5, respectively. The low values in the low-level evaluation set is explained by the inclusion of a parasitic organism into the test set, which has a reduced genome.

Computational performance

The computational time for the preliminary preparation steps, including HMM searches and alignments, increased linearly with each additional genome ($t = 55.597$, df = 63, $p < 0.001^{***}$, $R^2 = 0.98$) and the best fit line for this relationship exhibited a slope of 16.9 s/genome. The total runtime was also found to be significantly correlated with a linear model ($t = 45.9$, df = 63, $p < 0.001^{***}$, $R^2 = 0.97$) and this relationship exhibited a slope of 709.9 s/genome, although non-linear increases were observable for low-genome numbers (Fig. 5). This may be due to general RAxML initialization steps that are independent of data amount and are thus proportionally overrepresented with the low genome number analyses. In general, it can be assumed that the tree-building step, not the bcgTree specific tasks, consumes the largest fraction of the runtime.

In the current setup, a phylogenetic tree from core genomes of 50 organisms can be calculated on a standard desktop computer in less than 24 h. For larger analyses, the runtime can be decreased by using alternative tree calculation software or the high-performance computing variant ExaML that parallelizes tasks on different computer nodes.

Comparison with existing tools

The challenge of comparing whole genomes of bacteria has been undertaken through different approaches. One approach is to limit the scope to very closely related species to have large sets of orthologous genes. This approach is used by ATGC (Novichkov et al. 2009), which provides pre-calculated alignments for whole genomes

Fig. 4. Accuracy errors according to the quartet distance between trees of variable numbers of genes (1–108) and the trees based on the complete set. For each number of genes, 10 replicates were performed. For visualization purposes, the number of genes was rounded to increments of five. In addition, the quartet distance of the 16S rDNA tree is shown.

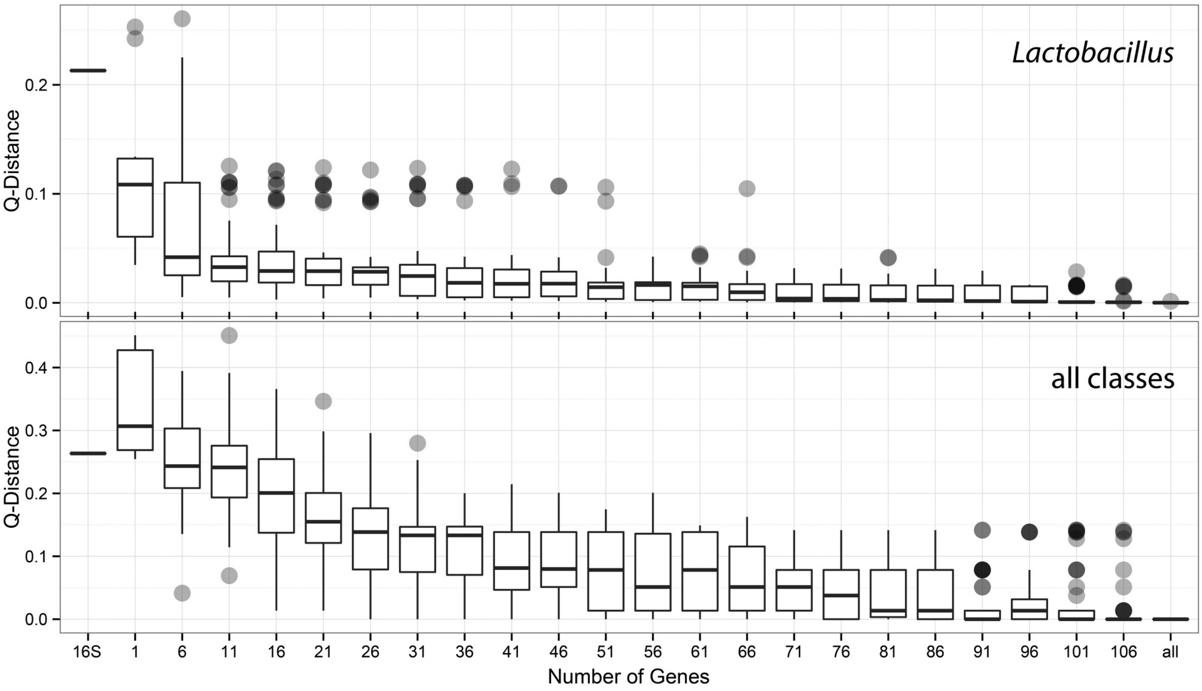
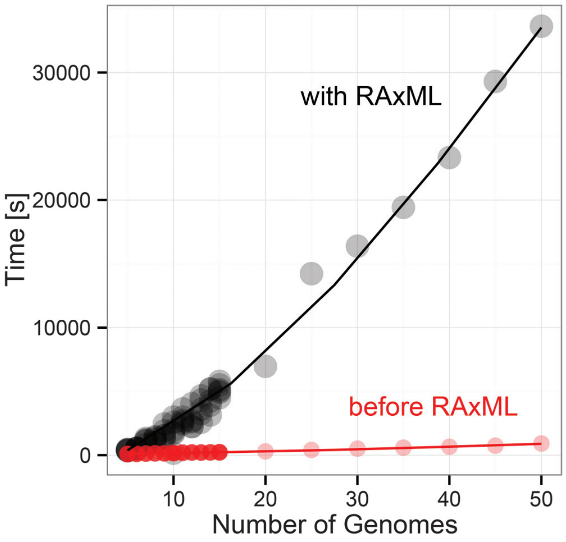


Fig. 5. Runtime of bcgTree for varying numbers of genomes in the tree calculation process with or without RAxML. [Colour online.]

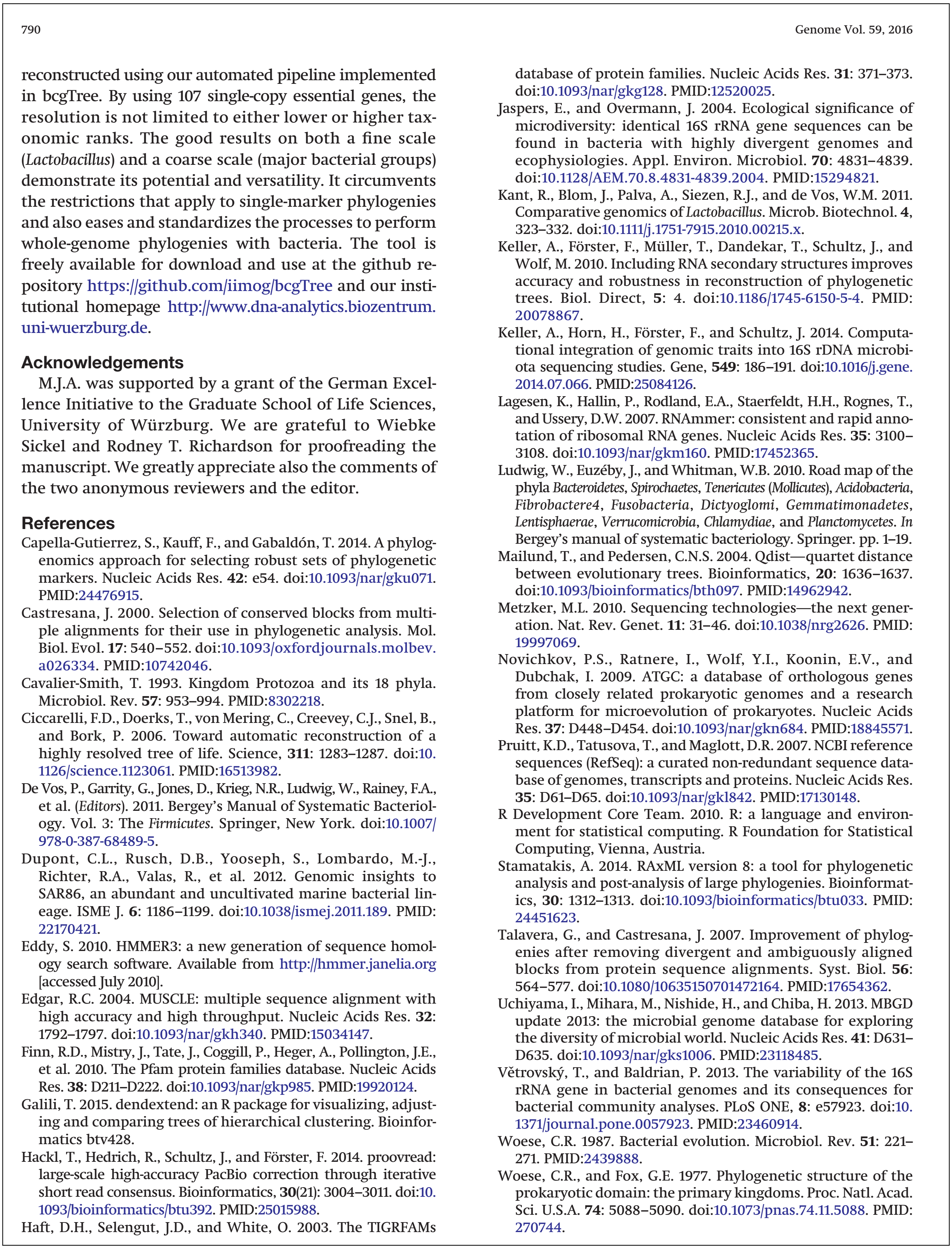


from NCBI RefSeq (Pruitt et al. 2007). This way, a large amount of information is available for each cluster (e.g., 1500 orthologous genes for *Lactobacillus*), providing a solid basis for micro-evolutionary analyses. Yet, only a very limited number of taxa can be included in ATGC analyses, for example only 11 *Lactobacillus* genomes. Also, ATGC is not designed to do analyses across clusters. The 11 *Lactobacillus* genomes are split over four clusters, so a comparison across the whole genus or with closely related genera is not practical. Further, including user-provided datasets in ATGC is not possible. In summary, ATGC can help answering micro-evolutionary questions, but it is limited for broader phylogenetic research questions.

Another approach is to use alignment-free methods using composition vectors, as implemented in CVTree (Zuo and Hao 2015). By dropping the alignment step many potential bioinformatic challenges (like length-hypervariable genes) can be avoided and the distances between genomes can be rapidly calculated. However, whilst more sequences are included (typically all proteins of a genome), the position information of each amino acid is dropped and thus a great amount of information ignored. As a consequence, the overall information content is quite different to the approach described in this article and might yield different results. We suggest using both the alignment-free and our alignment-based approach together for phylogenetic studies on whole genomes. Both methods are valid and may provide complementary and supportive viewpoints on bacterial phylogenies.

Conclusions

As demonstrated by the case study and evaluation, bacterial phylogenies can be accurately and robustly



reconstructed using our automated pipeline implemented in bcgTree. By using 107 single-copy essential genes, the resolution is not limited to either lower or higher taxonomic ranks. The good results on both a fine scale (*Lactobacillus*) and a coarse scale (major bacterial groups) demonstrate its potential and versatility. It circumvents the restrictions that apply to single-marker phylogenies and also eases and standardizes the processes to perform whole-genome phylogenies with bacteria. The tool is freely available for download and use at the github repository https://github.com/iimog/bcgTree and our institutional homepage http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

Acknowledgements

M.J.A. was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. We are grateful to Wiebke Sickel and Rodney T. Richardson for proofreading the manuscript. We greatly appreciate also the comments of the two anonymous reviewers and the editor.

References

Capella-Gutierrez, S., Kauff, F., and Gabaldón, T. 2014. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res. **42**: e54. doi:10.1093/nar/gku071. PMID:24476915.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**: 540–552. doi:10.1093/oxfordjournals.molbev.a026334. PMID:10742046.

Cavalier-Smith, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**: 953–994. PMID:8302218.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, **311**: 1283–1287. doi:10.1126/science.1123061. PMID:16513982.

De Vos, P., Garrity, G., Jones, D., Krieg, N.R., Ludwig, W., Rainey, F.A., et al. (*Editors*). 2011. Bergey's Manual of Systematic Bacteriology. Vol. 3: The *Firmicutes*. Springer, New York. doi:10.1007/978-0-387-68489-5.

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Richter, R.A., Valas, R., et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. **6**: 1186–1199. doi:10.1038/ismej.2011.189. PMID: 22170421.

Eddy, S. 2010. HMMER3: a new generation of sequence homology search software. Available from http://hmmer.janelia.org [accessed July 2010].

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**: 1792–1797. doi:10.1093/nar/gkh340. PMID:15034147.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., et al. 2010. The Pfam protein families database. Nucleic Acids Res. **38**: D211–D222. doi:10.1093/nar/gkp985. PMID:19920124.

Galili, T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics btv428.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, **30**(21): 3004–3011. doi:10.1093/bioinformatics/btu392. PMID:25015988.

Haft, D.H., Selengut, J.D., and White, O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. **31**: 371–373. doi:10.1093/nar/gkg128. PMID:12520025.

Jaspers, E., and Overmann, J. 2004. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl. Environ. Microbiol. **70**: 4831–4839. doi:10.1128/AEM.70.8.4831-4839.2004. PMID:15294821.

Kant, R., Blom, J., Palva, A., Siezen, R.J., and de Vos, W.M. 2011. Comparative genomics of *Lactobacillus*. Microb. Biotechnol. **4**, 323–332. doi:10.1111/j.1751-7915.2010.00215.x.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., and Wolf, M. 2010. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. Biol. Direct, **5**: 4. doi:10.1186/1745-6150-5-4. PMID: 20078867.

Keller, A., Horn, H., Förster, F., and Schultz, J. 2014. Computational integration of genomic traits into 16S rDNA microbiota sequencing studies. Gene, **549**: 186–191. doi:10.1016/j.gene.2014.07.066. PMID:25084126.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35**: 3100–3108. doi:10.1093/nar/gkm160. PMID:17452365.

Ludwig, W., Euzéby, J., and Whitman, W.B. 2010. Road map of the phyla *Bacteroidetes*, *Spirochaetes*, *Tenericutes* (*Mollicutes*), *Acidobacteria*, *Fibrobactere4*, *Fusobacteria*, *Dictyoglomi*, *Gemmatimonadetes*, *Lentisphaerae*, *Verrucomicrobia*, *Chlamydiae*, and *Planctomycetes*. *In* Bergey's manual of systematic bacteriology. Springer. pp. 1–19.

Mailund, T., and Pedersen, C.N.S. 2004. Qdist—quartet distance between evolutionary trees. Bioinformatics, **20**: 1636–1637. doi:10.1093/bioinformatics/bth097. PMID:14962942.

Metzker, M.L. 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. **11**: 31–46. doi:10.1038/nrg2626. PMID: 19997069.

Novichkov, P.S., Ratnere, I., Wolf, Y.I., Koonin, E.V., and Dubchak, I. 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res. **37**: D448–D454. doi:10.1093/nar/gkn684. PMID:18845571.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**: D61–D65. doi:10.1093/nar/gkl842. PMID:17130148.

R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, **30**: 1312–1313. doi:10.1093/bioinformatics/btu033. PMID: 24451623.

Talavera, G., and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. **56**: 564–577. doi:10.1080/10635150701472164. PMID:17654362.

Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. 2013. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucleic Acids Res. **41**: D631–D635. doi:10.1093/nar/gks1006. PMID:23118485.

Větrovský, T., and Baldrian, P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE, **8**: e57923. doi:10.1371/journal.pone.0057923. PMID:23460914.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271. PMID:2439888.

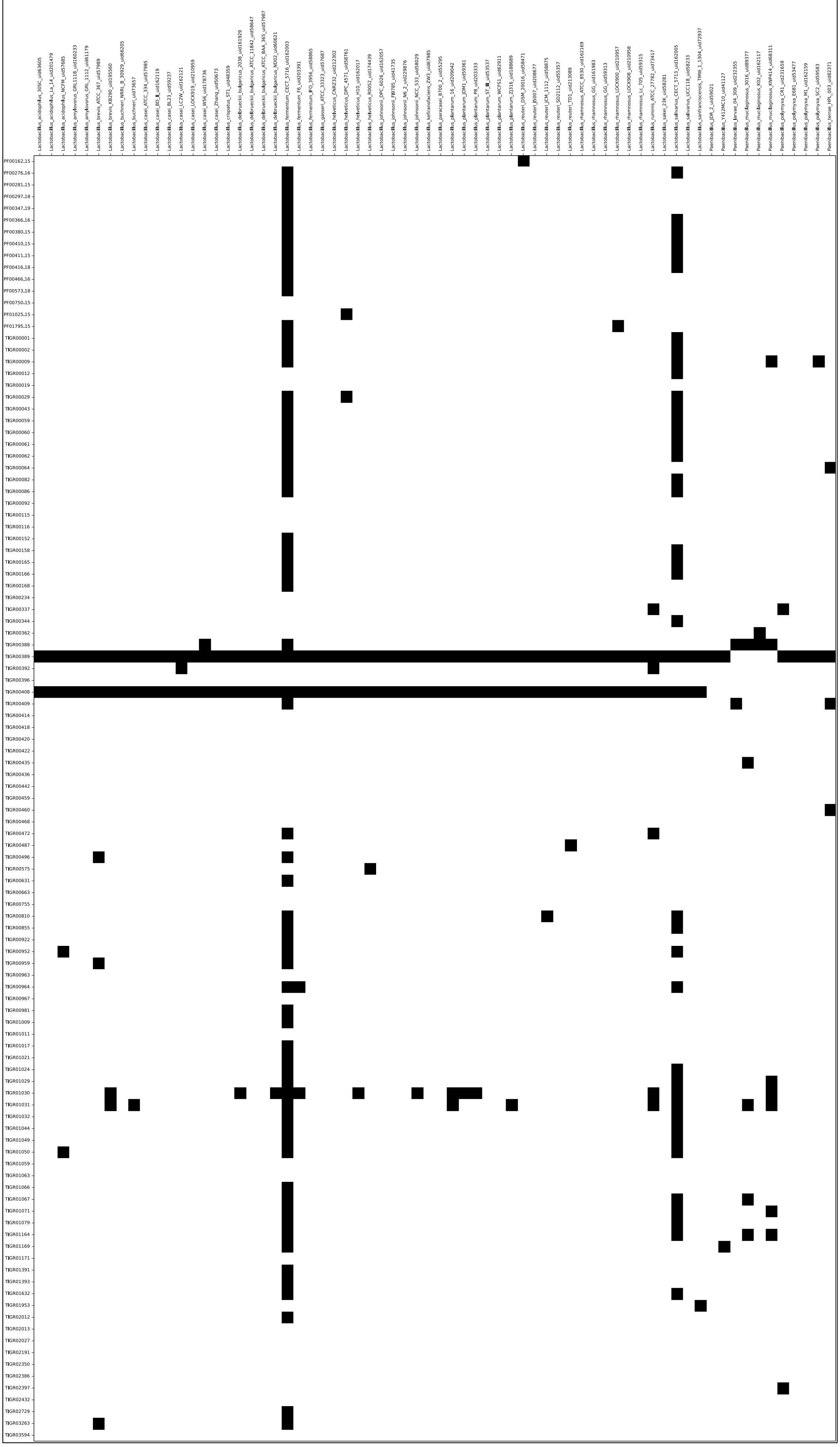
Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. **74**: 5088–5090. doi:10.1073/pnas.74.11.5088. PMID: 270744.

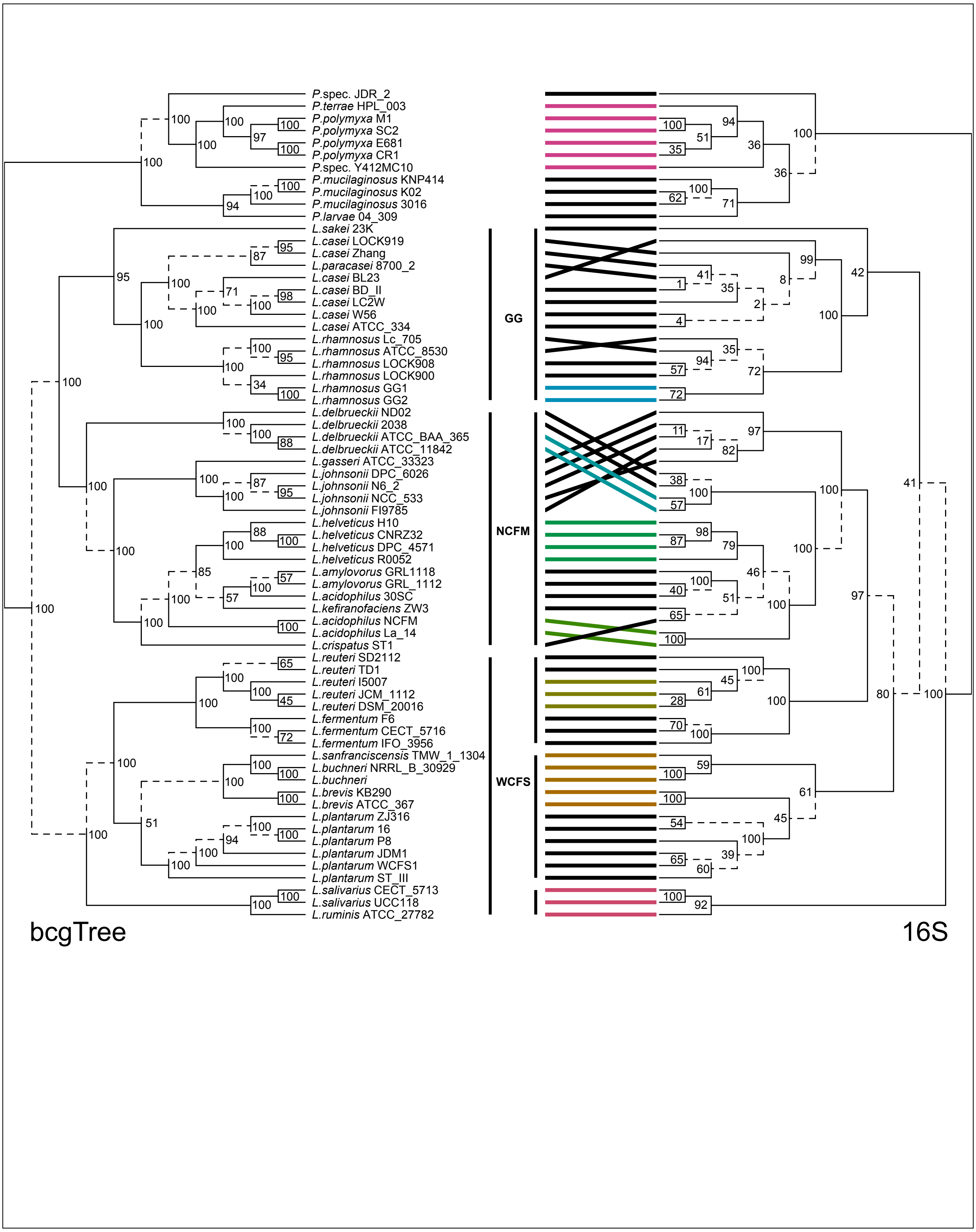
Wu, D., Jospin, G., and Eisen, J.A. 2013. Systematic identification of gene families for use as “markers” for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE, **8**: e77033. doi:10.1371/journal.pone.0077033. PMID:24146954.

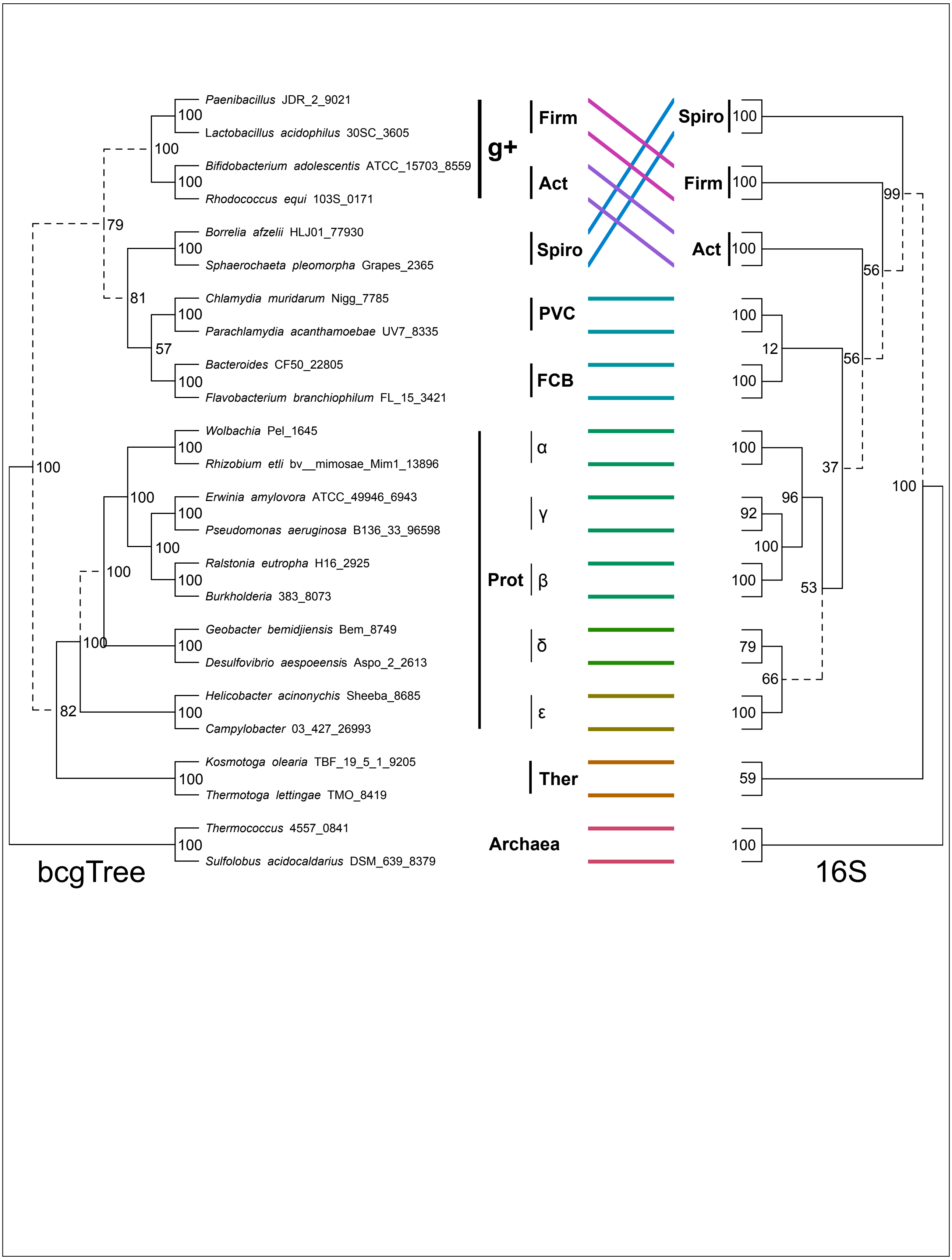
Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., et al. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. **31**: 241–250. doi:10.1016/j.syapm.2008.07.001. PMID:18692976.

Zuo, G., and Hao, B. 2015. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics Proteom. Bioinform. **13**: 321–331. doi:10.1016/j.gpb.2015.08.004.

Supplementary File 1: presence/absence matrix of essential core genes found for the *Lactobacillus* dataset. Rows are genes and columns are genomes.

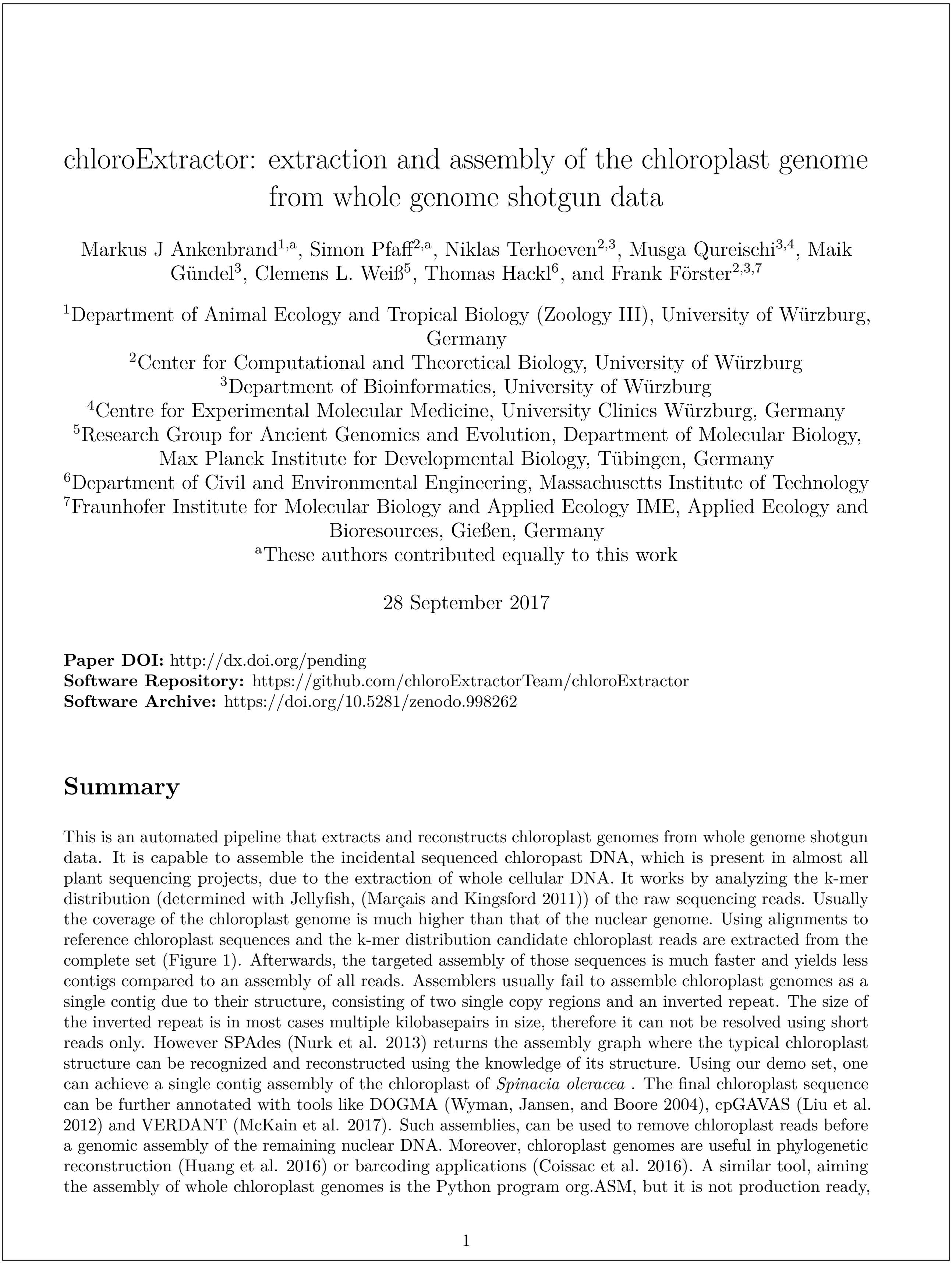



P.spec. JDR_2
P.terrae HPL_003
P.polymyxa M1
P.polymyxa SC2
P.polymyxa E681
P.polymyxa CR1
P.spec. Y412MC10
P.mucilaginosus KNP414
P.mucilaginosus K02
P.mucilaginosus 3016
P.larvae 04_309
L.sakei 23K
L.casei LOCK919
L.casei Zhang
L.paracasei 8700_2
L.casei BL23
L.casei BD_II
L.casei LC2W
L.casei W56
L.casei ATCC_334
L.rhamnosus Lc_705
L.rhamnosus ATCC_8530
L.rhamnosus LOCK908
L.rhamnosus LOCK900
L.rhamnosus GG1
L.rhamnosus GG2
L.delbrueckii ND02
L.delbrueckii 2038
L.delbrueckii ATCC_BAA_365
L.delbrueckii ATCC_11842
L.gasseri ATCC_33323
L.johnsonii DPC_6026
L.johnsonii N6_2
L.johnsonii NCC_533
L.johnsonii FI9785
L.helveticus H10
L.helveticus CNRZ32
L.helveticus DPC_4571
L.helveticus R0052
L.amylovorus GRL1118
L.amylovorus GRL_1112
L.acidophilus 30SC
L.kefiranofaciens ZW3
L.acidophilus NCFM
L.acidophilus La_14
L.crispatus ST1
L.reuteri SD2112
L.reuteri TD1
L.reuteri I5007
L.reuteri JCM_1112
L.reuteri DSM_20016
L.fermentum F6
L.fermentum CECT_5716
L.fermentum IFO_3956
L.sanfranciscensis TMW_1_1304
L.buchneri NRRL_B_30929
L.buchneri
L.brevis KB290
L.brevis ATCC_367
L.plantarum ZJ316
L.plantarum 16
L.plantarum P8
L.plantarum JDM1
L.plantarum WCFS1
L.plantarum ST_III
L.salivarius CECT_5713
L.salivarius UCC118
L.ruminis ATCC_27782
GG
NCFM
WCFS
bcgTree
16S


Paenibacillus JDR_2_9021
Lactobacillus acidophilus 30SC_3605
Bifidobacterium adolescentis ATCC_15703_8559
Rhodococcus equi 103S_0171
Borrelia afzelii HLJ01_77930
Sphaerochaeta pleomorpha Grapes_2365
Chlamydia muridarum Nigg_7785
Parachlamydia acanthamoebae UV7_8335
Bacteroides CF50_22805
Flavobacterium branchiophilum FL_15_3421
Wolbachia Pel_1645
Rhizobium etli bv__mimosae_Mim1_13896
Erwinia amylovora ATCC_49946_6943
Pseudomonas aeruginosa B136_33_96598
Ralstonia eutropha H16_2925
Burkholderia 383_8073
Geobacter bemidjiensis Bem_8749
Desulfovibrio aespoeensis Aspo_2_2613
Helicobacter acinonychis Sheeba_8685
Campylobacter 03_427_26993
Kosmotoga olearia TBF_19_5_1_9205
Thermotoga lettingae TMO_8419
Thermococcus 4557_0841
Sulfolobus acidocaldarius DSM_639_8379
100
79
81
57
82
bcgTree
g+
Firm
Act
Spiro
PVC
FCB
Prot
α
γ
β
δ
ε
Ther
Archaea
Spiro
Firm
Act
99
56
12
37
96
92
53
79
66
59
16S

5.9 chloroextractor: extraction and assembly of the chloroplast genome from whole genome shotgun data

– submitted to *Journal for Open Source Software* –



chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data

Markus J Ankenbrand[1,a], Simon Pfaff[2,a], Niklas Terhoeven[2,3], Musga Qureischi[3,4], Maik Gündel[3], Clemens L. Weiß[5], Thomas Hackl[6], and Frank Förster[2,3,7]

[1]Department of Animal Ecology and Tropical Biology (Zoology III), University of Würzburg, Germany
[2]Center for Computational and Theoretical Biology, University of Würzburg
[3]Department of Bioinformatics, University of Würzburg
[4]Centre for Experimental Molecular Medicine, University Clinics Würzburg, Germany
[5]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany
[6]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
[7]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Applied Ecology and Bioresources, Gießen, Germany
[a]These authors contributed equally to this work

28 September 2017

Paper DOI: http://dx.doi.org/pending
Software Repository: https://github.com/chloroExtractorTeam/chloroExtractor
Software Archive: https://doi.org/10.5281/zenodo.998262

Summary

This is an automated pipeline that extracts and reconstructs chloroplast genomes from whole genome shotgun data. It is capable to assemble the incidental sequenced chloropast DNA, which is present in almost all plant sequencing projects, due to the extraction of whole cellular DNA. It works by analyzing the k-mer distribution (determined with Jellyfish, (Marçais and Kingsford 2011)) of the raw sequencing reads. Usually the coverage of the chloroplast genome is much higher than that of the nuclear genome. Using alignments to reference chloroplast sequences and the k-mer distribution candidate chloroplast reads are extracted from the complete set (Figure 1). Afterwards, the targeted assembly of those sequences is much faster and yields less contigs compared to an assembly of all reads. Assemblers usually fail to assemble chloroplast genomes as a single contig due to their structure, consisting of two single copy regions and an inverted repeat. The size of the inverted repeat is in most cases multiple kilobasepairs in size, therefore it can not be resolved using short reads only. However SPAdes (Nurk et al. 2013) returns the assembly graph where the typical chloroplast structure can be recognized and reconstructed using the knowledge of its structure. Using our demo set, one can achieve a single contig assembly of the chloroplast of *Spinacia oleracea* . The final chloroplast sequence can be further annotated with tools like DOGMA (Wyman, Jansen, and Boore 2004), cpGAVAS (Liu et al. 2012) and VERDANT (McKain et al. 2017). Such assemblies, can be used to remove chloroplast reads before a genomic assembly of the remaining nuclear DNA. Moreover, chloroplast genomes are useful in phylogenetic reconstruction (Huang et al. 2016) or barcoding applications (Coissac et al. 2016). A similar tool, aiming the assembly of whole chloroplast genomes is the Python program org.ASM, but it is not production ready,

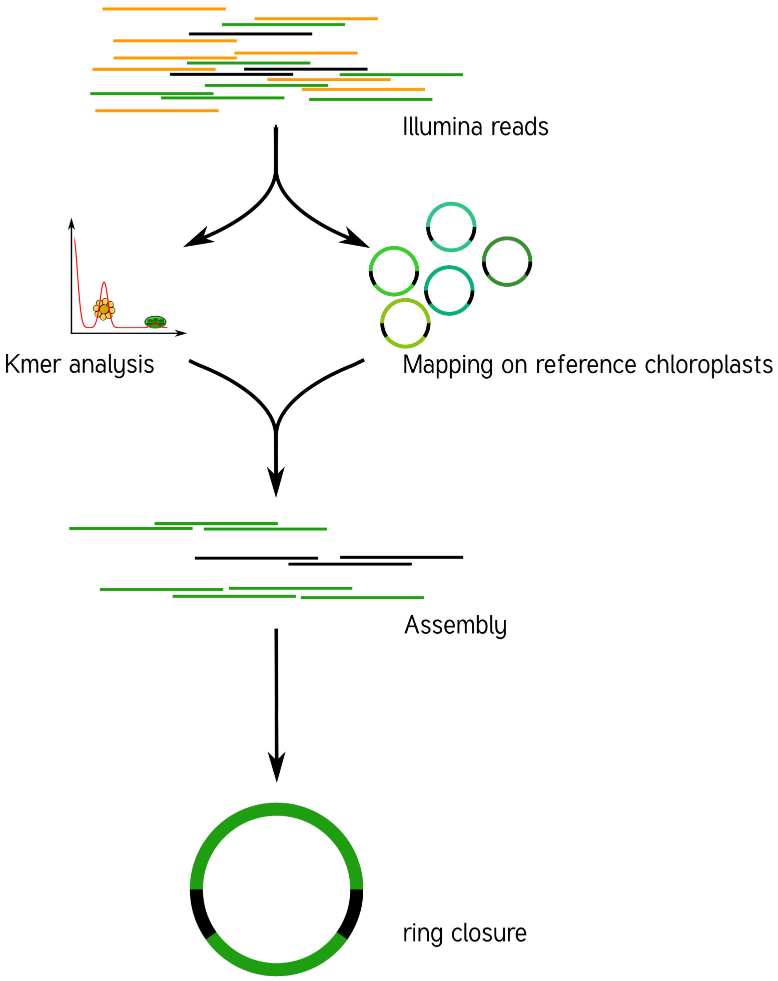
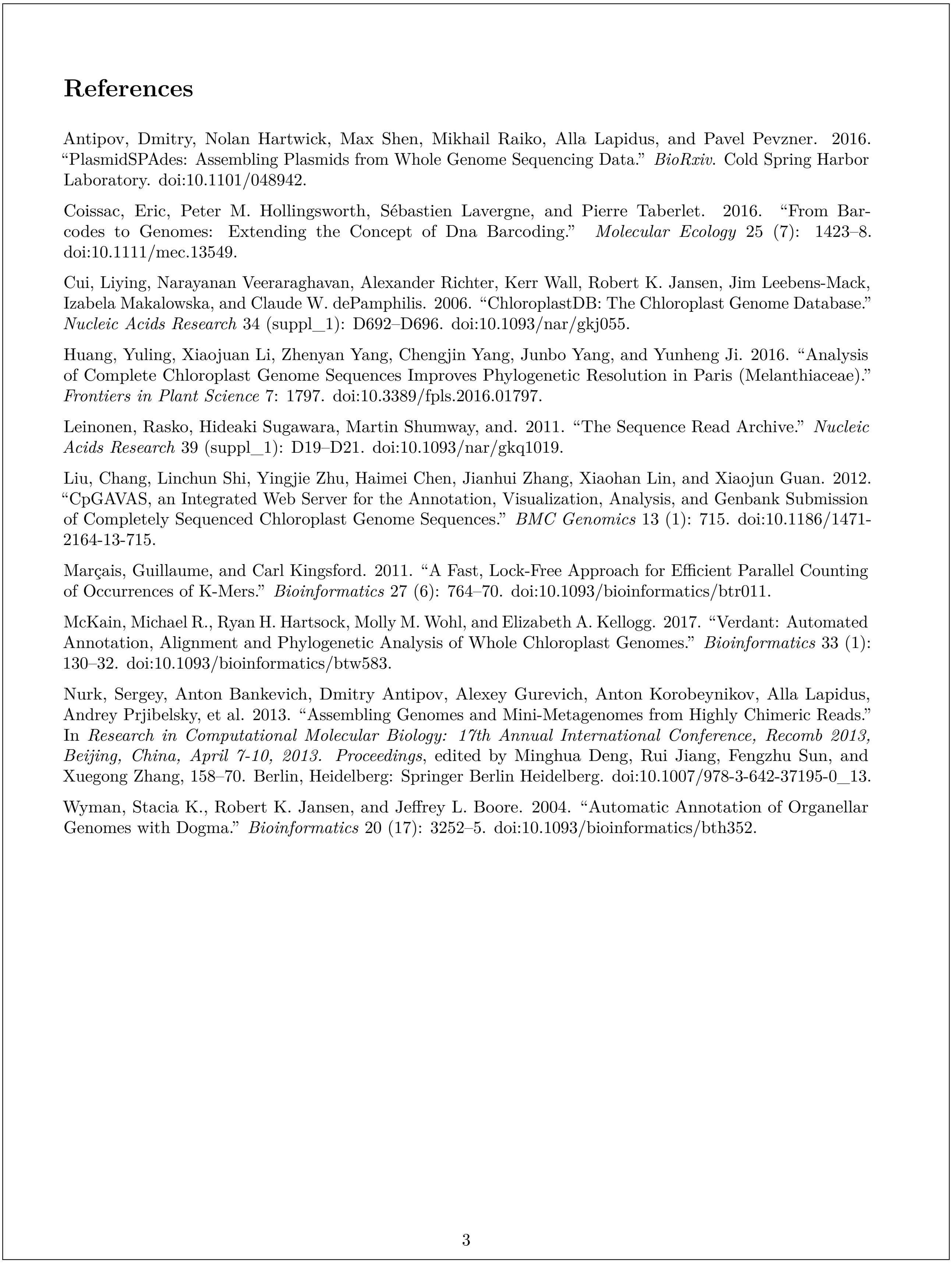


Figure 1: Schematic workflow of chloroExtractor.

yet. Also plasmid SPAdes (Antipov et al. 2016) could possibly be used for this purpose although it is not intended for it. In the future, we plan to use our chloroExtractor to screen NCBI's Sequence Read Archive (Leinonen et al. 2011) for chloroplast genomes in public sequencing datasets that are not yet available in chloroplast databases, eg. chloroDB (Cui et al. 2006) to broaden our knowledge about chloroplasts.

Acknowledgements

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.



References

Antipov, Dmitry, Nolan Hartwick, Max Shen, Mikhail Raiko, Alla Lapidus, and Pavel Pevzner. 2016. “PlasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data.” *BioRxiv.* Cold Spring Harbor Laboratory. doi:10.1101/048942.

Coissac, Eric, Peter M. Hollingsworth, Sébastien Lavergne, and Pierre Taberlet. 2016. “From Barcodes to Genomes: Extending the Concept of Dna Barcoding.” *Molecular Ecology* 25 (7): 1423–8. doi:10.1111/mec.13549.

Cui, Liying, Narayanan Veeraraghavan, Alexander Richter, Kerr Wall, Robert K. Jansen, Jim Leebens-Mack, Izabela Makalowska, and Claude W. dePamphilis. 2006. “ChloroplastDB: The Chloroplast Genome Database.” *Nucleic Acids Research* 34 (suppl_1): D692–D696. doi:10.1093/nar/gkj055.

Huang, Yuling, Xiaojuan Li, Zhenyan Yang, Chengjin Yang, Junbo Yang, and Yunheng Ji. 2016. “Analysis of Complete Chloroplast Genome Sequences Improves Phylogenetic Resolution in Paris (Melanthiaceae).” *Frontiers in Plant Science* 7: 1797. doi:10.3389/fpls.2016.01797.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and. 2011. “The Sequence Read Archive.” *Nucleic Acids Research* 39 (suppl_1): D19–D21. doi:10.1093/nar/gkq1019.

Liu, Chang, Linchun Shi, Yingjie Zhu, Haimei Chen, Jianhui Zhang, Xiaohan Lin, and Xiaojun Guan. 2012. “CpGAVAS, an Integrated Web Server for the Annotation, Visualization, Analysis, and Genbank Submission of Completely Sequenced Chloroplast Genome Sequences.” *BMC Genomics* 13 (1): 715. doi:10.1186/1471-2164-13-715.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70. doi:10.1093/bioinformatics/btr011.

McKain, Michael R., Ryan H. Hartsock, Molly M. Wohl, and Elizabeth A. Kellogg. 2017. “Verdant: Automated Annotation, Alignment and Phylogenetic Analysis of Whole Chloroplast Genomes.” *Bioinformatics* 33 (1): 130–32. doi:10.1093/bioinformatics/btw583.

Nurk, Sergey, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, et al. 2013. “Assembling Genomes and Mini-Metagenomes from Highly Chimeric Reads.” In *Research in Computational Molecular Biology: 17th Annual International Conference, Recomb 2013, Beijing, China, April 7-10, 2013. Proceedings*, edited by Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, 158–70. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-37195-0_13.

Wyman, Stacia K., Robert K. Jansen, and Jeffrey L. Boore. 2004. “Automatic Annotation of Organellar Genomes with Dogma.” *Bioinformatics* 20 (17): 3252–5. doi:10.1093/bioinformatics/bth352.

5.10 ALITV-INTERACTIVE VISUALIZATION OF WHOLE GENOME COMPARISONS

– published in *PeerJ Computer Science* –

PeerJ Computer Science

AliTV—interactive visualization of whole genome comparisons

Markus J. Ankenbrand[1,*], Sonja Hohlfeld[1,2,*], Thomas Hackl[2,3] and Frank Förster[2,4]

[1] Department of Animal Ecology and Tropical Biology, Julius Maximilian University, Würzburg, Germany
[2] Department for Bioinformatics, Julius Maximilian University, Würzburg, Germany
[3] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[4] Center for Computational and Theoretical Biology, Julius Maximilian University, Würzburg, Germany
[*] These authors contributed equally to this work.

ABSTRACT

Whole genome alignments and comparative analysis are key methods in the quest of unraveling the dynamics of genome evolution. Interactive visualization and exploration of the generated alignments, annotations, and phylogenetic data are important steps in the interpretation of the initial results. Limitations of existing software inspired us to develop our new tool AliTV, which provides interactive visualization of whole genome alignments. AliTV reads multiple whole genome alignments or automatically generates alignments from the provided data. Optional feature annotations and phylogenetic information are supported. The user-friendly, web-browser based and highly customizable interface allows rapid exploration and manipulation of the visualized data as well as the export of publication-ready high-quality figures. AliTV is freely available at https://github.com/AliTVTeam/AliTV.

Subjects Bioinformatics, Computational Biology
Keywords Comparative genomics, Alignment, Visualization

Submitted 22 April 2017
Accepted 25 April 2017
Published 12 June 2017

Corresponding author
Frank Förster,
frank.foerster@uni-wuerzburg.de,
foersterfrank@gmx.de

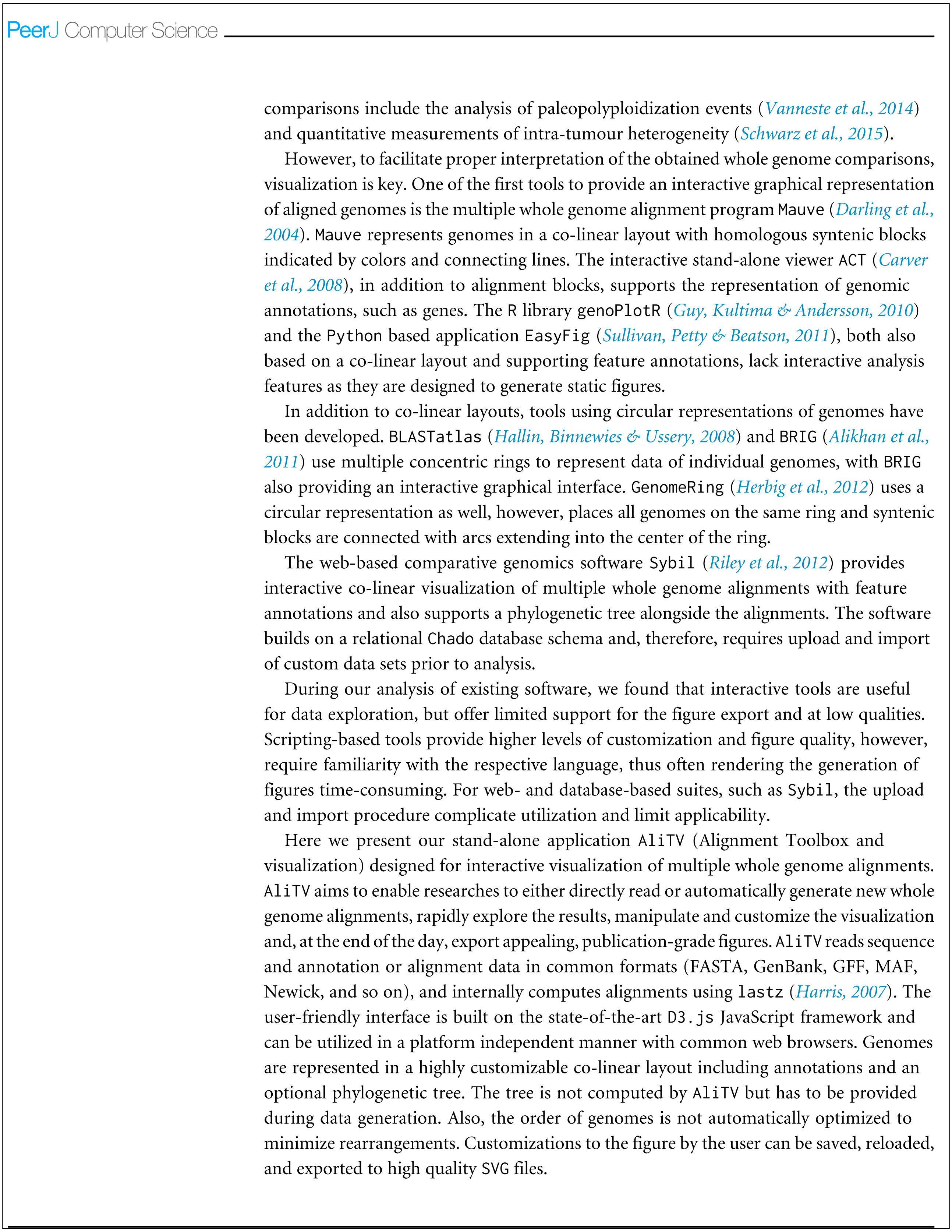
Academic editor
Shawn Gomez

Additional Information and Declarations can be found on page 7

DOI 10.7717/peerj-cs.116

INTRODUCTION

Advances in short- and long-read sequencing and assembly over the last decade (*Salzberg et al., 2011*; *Chin et al., 2013*; *Hackl et al., 2014*) have made whole genome sequencing a routine task for biologists in various fields. Public sequence databases already contain several thousand of draft and finished genomes (*Benson et al., 2013*), with many more on the way (*Pagani et al., 2012*). In particular, high throughput sequencing projects of pathogen strains related to recent outbreaks (*Rasko et al., 2011*), and large-scale ecological studies targeting microbial communities and pan genomes of populations using metagenome and single cell sequencing approaches contribute in this process (*Turnbaugh et al., 2007*; *Kashtan et al., 2014*). These rich data sets can be explored for large-scale evolutionary processes using comparative genomics and whole genome alignments, revealing genomic recombinations (*Didelot, Méric & Falush, 2012*; *Namouchi et al., 2012*; *Yahara et al., 2014*), islands and horizontal gene transfer (*Avrani et al., 2011*; *Coleman et al., 2006*; *Langille, Hsiao & Brinkman, 2008*) as well as the often related dynamics of mobile or endogenous viral elements (*Fischer, 2015*; *Touchon & Rocha, 2007*). Other applications of whole genome



comparisons include the analysis of paleopolyploidization events (Vanneste et al., 2014) and quantitative measurements of intra-tumour heterogeneity (Schwarz et al., 2015).

However, to facilitate proper interpretation of the obtained whole genome comparisons, visualization is key. One of the first tools to provide an interactive graphical representation of aligned genomes is the multiple whole genome alignment program Mauve (Darling et al., 2004). Mauve represents genomes in a co-linear layout with homologous syntenic blocks indicated by colors and connecting lines. The interactive stand-alone viewer ACT (Carver et al., 2008), in addition to alignment blocks, supports the representation of genomic annotations, such as genes. The R library genoPlotR (Guy, Kultima & Andersson, 2010) and the Python based application EasyFig (Sullivan, Petty & Beatson, 2011), both also based on a co-linear layout and supporting feature annotations, lack interactive analysis features as they are designed to generate static figures.

In addition to co-linear layouts, tools using circular representations of genomes have been developed. BLASTatlas (Hallin, Binnewies & Ussery, 2008) and BRIG (Alikhan et al., 2011) use multiple concentric rings to represent data of individual genomes, with BRIG also providing an interactive graphical interface. GenomeRing (Herbig et al., 2012) uses a circular representation as well, however, places all genomes on the same ring and syntenic blocks are connected with arcs extending into the center of the ring.

The web-based comparative genomics software Sybil (Riley et al., 2012) provides interactive co-linear visualization of multiple whole genome alignments with feature annotations and also supports a phylogenetic tree alongside the alignments. The software builds on a relational Chado database schema and, therefore, requires upload and import of custom data sets prior to analysis.

During our analysis of existing software, we found that interactive tools are useful for data exploration, but offer limited support for the figure export and at low qualities. Scripting-based tools provide higher levels of customization and figure quality, however, require familiarity with the respective language, thus often rendering the generation of figures time-consuming. For web- and database-based suites, such as Sybil, the upload and import procedure complicate utilization and limit applicability.

Here we present our stand-alone application AliTV (Alignment Toolbox and visualization) designed for interactive visualization of multiple whole genome alignments. AliTV aims to enable researches to either directly read or automatically generate new whole genome alignments, rapidly explore the results, manipulate and customize the visualization and, at the end of the day, export appealing, publication-grade figures. AliTV reads sequence and annotation or alignment data in common formats (FASTA, GenBank, GFF, MAF, Newick, and so on), and internally computes alignments using lastz (Harris, 2007). The user-friendly interface is built on the state-of-the-art D3.js JavaScript framework and can be utilized in a platform independent manner with common web browsers. Genomes are represented in a highly customizable co-linear layout including annotations and an optional phylogenetic tree. The tree is not computed by AliTV but has to be provided during data generation. Also, the order of genomes is not automatically optimized to minimize rearrangements. Customizations to the figure by the user can be saved, reloaded, and exported to high quality SVG files.

METHODS

Our tool AliTV is divided into two parts. The first non-interactive part is required for the generation of the input files for our interactive viewer. The second part represents that interactive viewer in the form of a SVG file embedded in a HTML5 website. The latest version of our code can be obtained from GitHub (https://github.com/AliTVTeam/AliTV). It is planned to adjust AliTV in order to integrate it into the BioJS registry (https://biojsnet.herokuapp.com/, *Corpas et al. (2014)*). The general design of AliTV assures, that AliTV runs on different hard- and software platforms, e.g., Linux, MacOSX, and Windows. The following sections describe those parts in more detail.

Data preparation

The data preparation is performed by a single Perl script named alitv.pl. This script uses a set of different Perl modules to import incoming data and generate valid JSON input data for our visualization engine described in the next paragraph. One of our aims is to support as many different input formats for sequence and annotation information as possible. Therefore, we used the well tested and broadly accepted BioPerl as basis for our modules (*Stajich et al., 2002*).

The script alitv.pl uses a YAML file to specify the different input files. Moreover, an easy-to-use-mode is available which requires only a couple of input files and generates the required YAML file on the fly. This generated YAML settings file might be used to reproduce AliTV results or can be used as starting point to alter configuration parameters.

During the preparation step, AliTV requires all-vs-all alignments of the complete sequence set. Those alignments are generated or user provided. The current version of alitv.pl requires lastz to generate all alignments in MAF format (*Harris, 2007*). Nevertheless, BioPerl supports a broad range of alignment formats. Therefore, other programs can easily be added to the list of supported alignment programs. Moreover, the ability to use existing alignments allows a huge time benefit, when AliTV parameters are changed to optimize the visualization via YAML settings file in a non-interactive manner. Thus future versions of alitv.pl will support caching of alignments based on checksums to avoid unnecessary recalculations.

The final result of our alitv.pl is a JSON file, which can be load into our interactive visualization page.

Interactive visualization

AliTV is implemented in JavaScript. Our code is documented using JSDoc 3 (version 3.4.0 http://usejsdoc.org/, 02.06.2016). AliTV generates a SVG which is presented within a browser using HTML5. A tutorial is available at https://alitv.readthedocs.io/en/latest/index.html.

To gain advanced application possibilities we use different libraries. The JavaScript library D3.js 3.5.17 (http://d3js.org/, 06.06.2016) provides a wide range of pre-built functions for calculating and drawing the interactive figure. In addition, AliTV employs JQuery 2.2.4 (https://jquery.com/, 06.06.2016) to ease access to several parts of the figure. This is helpful for hiding selected sequences, genes or links. JQueryUI 1.11.4 (https://jqueryui.com/, 06.06.2016) gives us the possibilities to add user-friendly

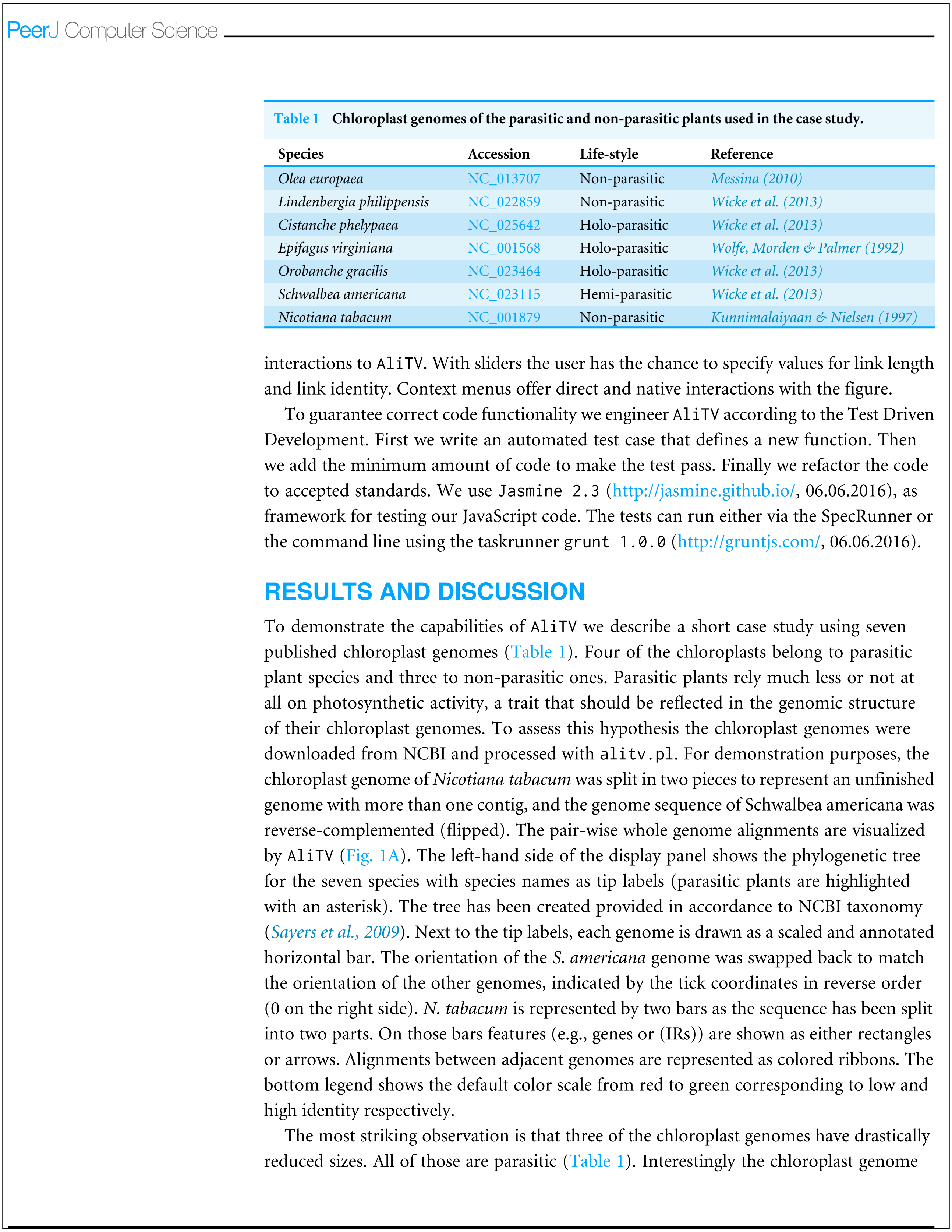


Table 1 Chloroplast genomes of the parasitic and non-parasitic plants used in the case study.

Species	Accession	Life-style	Reference
Olea europaea	NC_013707	Non-parasitic	*Messina (2010)*
Lindenbergia philippensis	NC_022859	Non-parasitic	*Wicke et al. (2013)*
Cistanche phelypaea	NC_025642	Holo-parasitic	*Wicke et al. (2013)*
Epifagus virginiana	NC_001568	Holo-parasitic	*Wolfe, Morden & Palmer (1992)*
Orobanche gracilis	NC_023464	Holo-parasitic	*Wicke et al. (2013)*
Schwalbea americana	NC_023115	Hemi-parasitic	*Wicke et al. (2013)*
Nicotiana tabacum	NC_001879	Non-parasitic	*Kunnimalaiyaan & Nielsen (1997)*

interactions to `AliTV`. With sliders the user has the chance to specify values for link length and link identity. Context menus offer direct and native interactions with the figure.

To guarantee correct code functionality we engineer `AliTV` according to the Test Driven Development. First we write an automated test case that defines a new function. Then we add the minimum amount of code to make the test pass. Finally we refactor the code to accepted standards. We use `Jasmine 2.3` (http://jasmine.github.io/, 06.06.2016), as framework for testing our JavaScript code. The tests can run either via the SpecRunner or the command line using the taskrunner `grunt 1.0.0` (http://gruntjs.com/, 06.06.2016).

RESULTS AND DISCUSSION

To demonstrate the capabilities of `AliTV` we describe a short case study using seven published chloroplast genomes (Table 1). Four of the chloroplasts belong to parasitic plant species and three to non-parasitic ones. Parasitic plants rely much less or not at all on photosynthetic activity, a trait that should be reflected in the genomic structure of their chloroplast genomes. To assess this hypothesis the chloroplast genomes were downloaded from NCBI and processed with `alitv.pl`. For demonstration purposes, the chloroplast genome of *Nicotiana tabacum* was split in two pieces to represent an unfinished genome with more than one contig, and the genome sequence of Schwalbea americana was reverse-complemented (flipped). The pair-wise whole genome alignments are visualized by `AliTV` (Fig. 1A). The left-hand side of the display panel shows the phylogenetic tree for the seven species with species names as tip labels (parasitic plants are highlighted with an asterisk). The tree has been created provided in accordance to NCBI taxonomy (*Sayers et al., 2009*). Next to the tip labels, each genome is drawn as a scaled and annotated horizontal bar. The orientation of the *S. americana* genome was swapped back to match the orientation of the other genomes, indicated by the tick coordinates in reverse order (0 on the right side). *N. tabacum* is represented by two bars as the sequence has been split into two parts. On those bars features (e.g., genes or (IRs)) are shown as either rectangles or arrows. Alignments between adjacent genomes are represented as colored ribbons. The bottom legend shows the default color scale from red to green corresponding to low and high identity respectively.

The most striking observation is that three of the chloroplast genomes have drastically reduced sizes. All of those are parasitic (Table 1). Interestingly the chloroplast genome

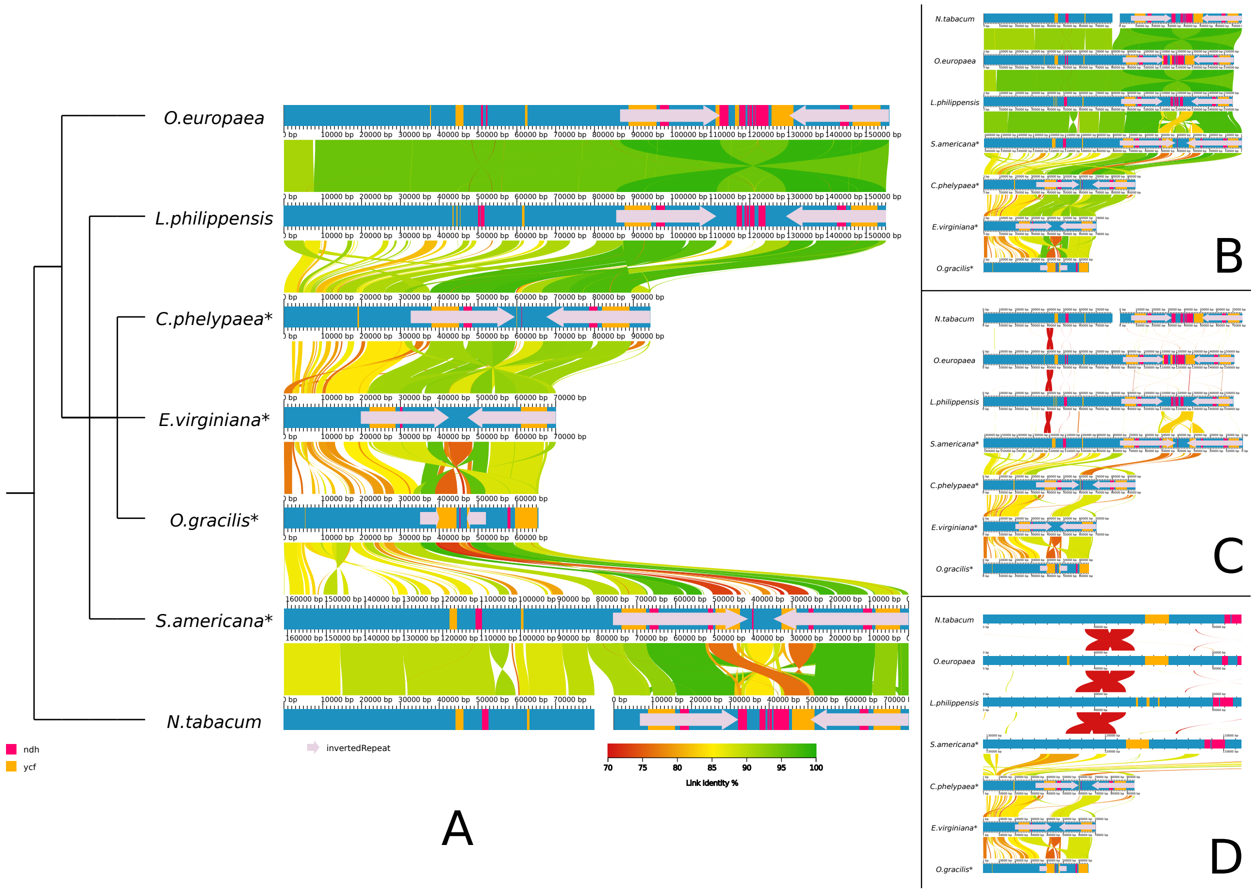
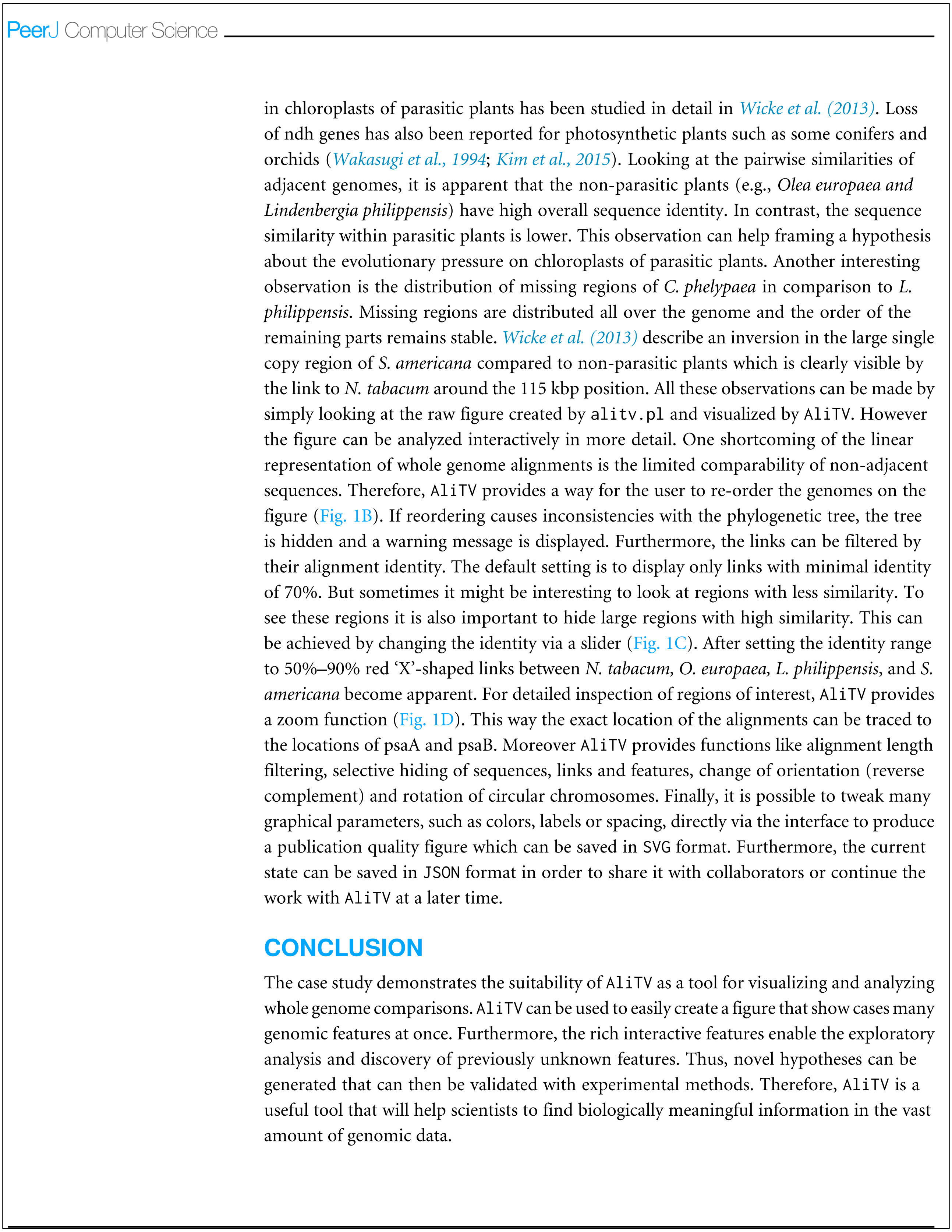


Figure 1 **Whole genome alignment of seven chloroplasts visualized by `AliTV`.** Species names were italicized and parasites marked with asterisks ex post. (A) Default layout with a phylogenetic tree on the left-hand side and genomes represented by co-linear horizontal bars on the right; genes and inverted repeats are displayed as rectangles and arrows, respectively; colored ribbons connect corresponding regions in the alignment. (B–D) Customized layouts: (B) reordered genomes, non-parasitic plants at the top and holo-parasitic plants at the bottom; (C) links filtered by identity (only those with 50%–90% identity are drawn); (D) zoom in on a potential segmental duplication (red 'X'-shaped links) in the top four genomes.

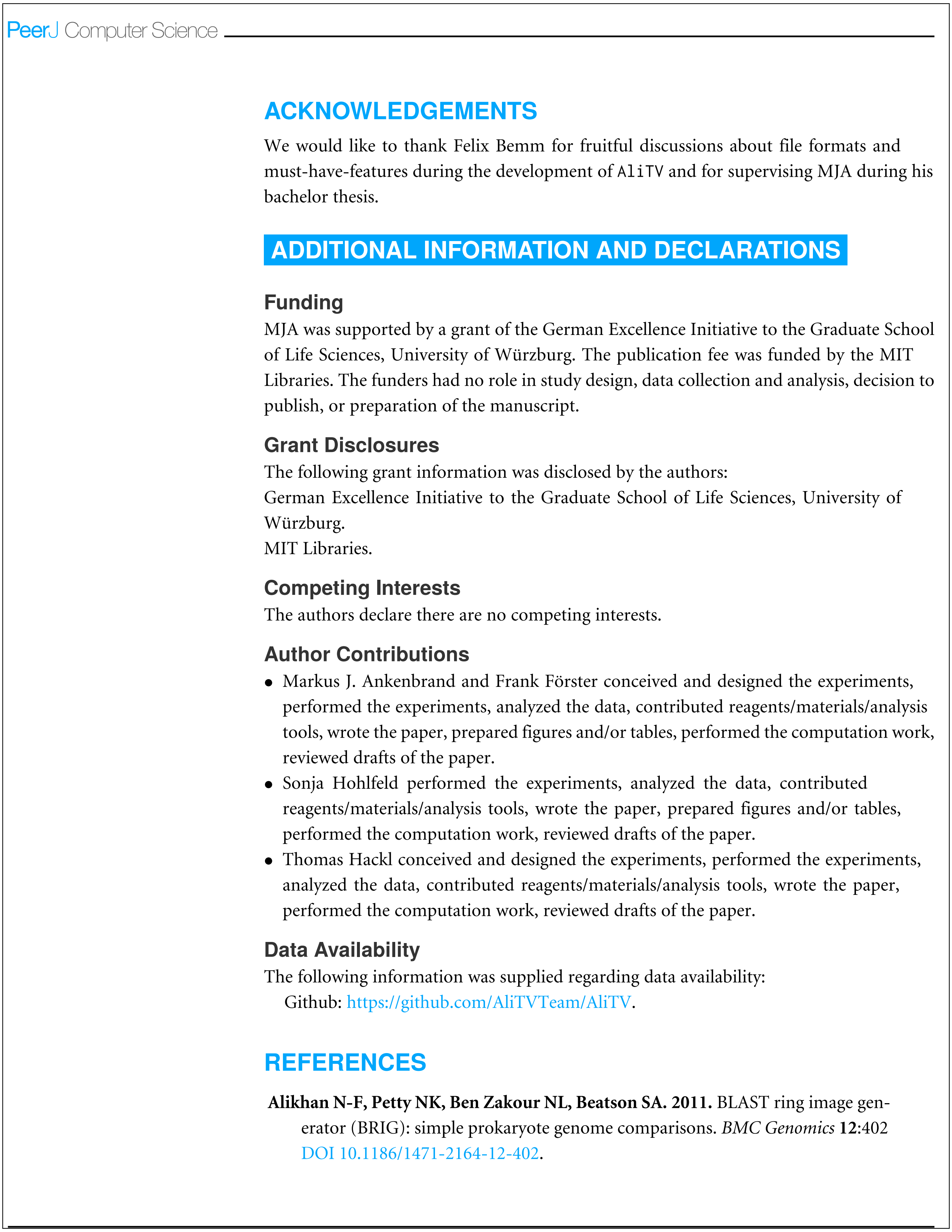
size of *S. americana* is similar to that of the non-parasitic plants. This can be explained by the life style of *S. americana* which is hemi-parasitic in contrast to the other parasitic plants which are holo-parasites. The features shown are the IR regions as arrows, the hypothetical chloroplast open reading frames as orange and the genes of the ndh family as pink rectangles. First, it can be seen that there is a big variation in size of the inverted repeats. While the IR of Orobanche gracilis is the shortest with roughly 5,000 bp, that of *S. americana* is the largest with roughly 35,000 bp. Second, there are less genes of the ndh family on *Cistanche phelypaea, Epifagus virginiana, O. gracilis*, and *S. americana*. Members of the ndh gene family encode subunits of the NADH dehydrogenase-like complex, which is involved in chlororespiration (*Martín & Sabater, 2010*). However, they are not required for plant growth under optimal conditions (*Burrows, 1998*). The absence of ndh genes



in chloroplasts of parasitic plants has been studied in detail in Wicke et al. (2013). Loss of ndh genes has also been reported for photosynthetic plants such as some conifers and orchids (Wakasugi et al., 1994; Kim et al., 2015). Looking at the pairwise similarities of adjacent genomes, it is apparent that the non-parasitic plants (e.g., *Olea europaea and Lindenbergia philippensis*) have high overall sequence identity. In contrast, the sequence similarity within parasitic plants is lower. This observation can help framing a hypothesis about the evolutionary pressure on chloroplasts of parasitic plants. Another interesting observation is the distribution of missing regions of *C. phelypaea* in comparison to *L. philippensis*. Missing regions are distributed all over the genome and the order of the remaining parts remains stable. Wicke et al. (2013) describe an inversion in the large single copy region of *S. americana* compared to non-parasitic plants which is clearly visible by the link to *N. tabacum* around the 115 kbp position. All these observations can be made by simply looking at the raw figure created by `alitv.pl` and visualized by `AliTV`. However the figure can be analyzed interactively in more detail. One shortcoming of the linear representation of whole genome alignments is the limited comparability of non-adjacent sequences. Therefore, `AliTV` provides a way for the user to re-order the genomes on the figure (Fig. 1B). If reordering causes inconsistencies with the phylogenetic tree, the tree is hidden and a warning message is displayed. Furthermore, the links can be filtered by their alignment identity. The default setting is to display only links with minimal identity of 70%. But sometimes it might be interesting to look at regions with less similarity. To see these regions it is also important to hide large regions with high similarity. This can be achieved by changing the identity via a slider (Fig. 1C). After setting the identity range to 50%–90% red 'X'-shaped links between *N. tabacum, O. europaea, L. philippensis*, and *S. americana* become apparent. For detailed inspection of regions of interest, `AliTV` provides a zoom function (Fig. 1D). This way the exact location of the alignments can be traced to the locations of psaA and psaB. Moreover `AliTV` provides functions like alignment length filtering, selective hiding of sequences, links and features, change of orientation (reverse complement) and rotation of circular chromosomes. Finally, it is possible to tweak many graphical parameters, such as colors, labels or spacing, directly via the interface to produce a publication quality figure which can be saved in `SVG` format. Furthermore, the current state can be saved in `JSON` format in order to share it with collaborators or continue the work with `AliTV` at a later time.

CONCLUSION

The case study demonstrates the suitability of `AliTV` as a tool for visualizing and analyzing whole genome comparisons. `AliTV` can be used to easily create a figure that show cases many genomic features at once. Furthermore, the rich interactive features enable the exploratory analysis and discovery of previously unknown features. Thus, novel hypotheses can be generated that can then be validated with experimental methods. Therefore, `AliTV` is a useful tool that will help scientists to find biologically meaningful information in the vast amount of genomic data.



ACKNOWLEDGEMENTS

We would like to thank Felix Bemm for fruitful discussions about file formats and must-have-features during the development of `AliTV` and for supervising MJA during his bachelor thesis.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. The publication fee was funded by the MIT Libraries. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.
MIT Libraries.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Markus J. Ankenbrand and Frank Förster conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Sonja Hohlfeld performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Thomas Hackl conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, performed the computation work, reviewed drafts of the paper.

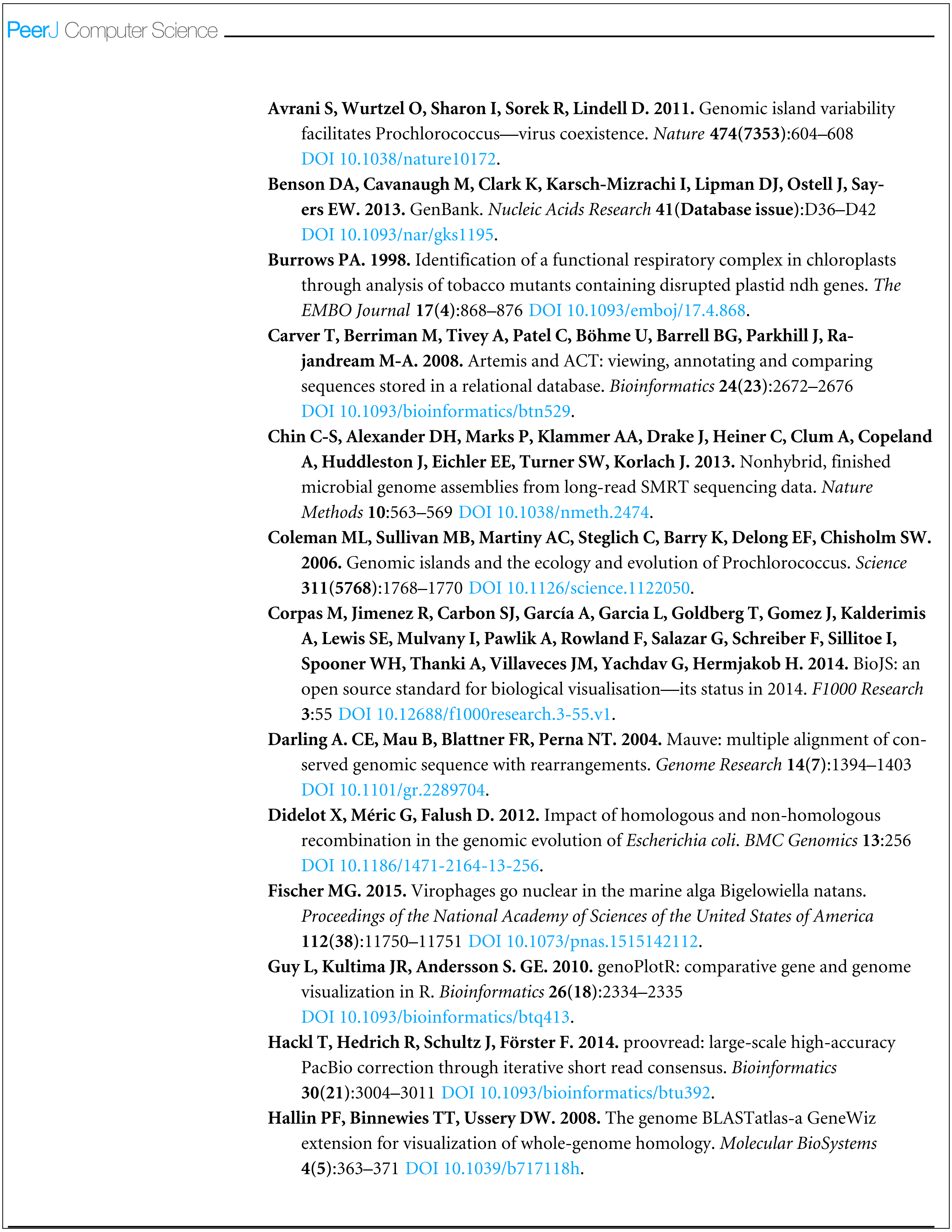
Data Availability

The following information was supplied regarding data availability:

Github: https://github.com/AliTVTeam/AliTV.

REFERENCES

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**:402 DOI 10.1186/1471-2164-12-402.



PeerJ Computer Science

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. 2011. Genomic island variability facilitates Prochlorococcus—virus coexistence. *Nature* **474(7353)**:604–608 DOI 10.1038/nature10172.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research* **41(Database issue)**:D36–D42 DOI 10.1093/nar/gks1195.

Burrows PA. 1998. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes. *The EMBO Journal* **17(4)**:868–876 DOI 10.1093/emboj/17.4.868.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M-A. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24(23)**:2672–2676 DOI 10.1093/bioinformatics/btn529.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**:563–569 DOI 10.1038/nmeth.2474.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311(5768)**:1768–1770 DOI 10.1126/science.1122050.

Corpas M, Jimenez R, Carbon SJ, García A, Garcia L, Goldberg T, Gomez J, Kalderimis A, Lewis SE, Mulvany I, Pawlik A, Rowland F, Salazar G, Schreiber F, Sillitoe I, Spooner WH, Thanki A, Villaveces JM, Yachdav G, Hermjakob H. 2014. BioJS: an open source standard for biological visualisation—its status in 2014. *F1000 Research* **3**:55 DOI 10.12688/f1000research.3-55.v1.

Darling A. CE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14(7)**:1394–1403 DOI 10.1101/gr.2289704.

Didelot X, Méric G, Falush D. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**:256 DOI 10.1186/1471-2164-13-256.

Fischer MG. 2015. Virophages go nuclear in the marine alga Bigelowiella natans. *Proceedings of the National Academy of Sciences of the United States of America* **112(38)**:11750–11751 DOI 10.1073/pnas.1515142112.

Guy L, Kultima JR, Andersson S. GE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26(18)**:2334–2335 DOI 10.1093/bioinformatics/btq413.

Hackl T, Hedrich R, Schultz J, Förster F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30(21)**:3004–3011 DOI 10.1093/bioinformatics/btu392.

Hallin PF, Binnewies TT, Ussery DW. 2008. The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. *Molecular BioSystems* **4(5)**:363–371 DOI 10.1039/b717118h.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University.

Herbig A, Jäger G, Battke F, Nieselt K. 2012. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics* **28(12)**:i7–i15 DOI 10.1093/bioinformatics/bts217.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science* **344(6182)**:416–420 DOI 10.1126/science.1248575.

Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim J-H. 2015. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *PLOS ONE* **10(11)**:e0142215 DOI 10.1371/journal.pone.0142215.

Kunnimalaiyaan M, Nielsen BL. 1997. Fine mapping of replication origins (ori A and ori B) in *Nicotiana tabacum* chloroplast DNA. *Nucleic Acids Research* **25(18)**:3681–3686 DOI 10.1093/nar/25.18.3681.

Langille M, Hsiao W, Brinkman F. 2008. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* **9**:329 DOI 10.1186/1471-2105-9-329.

Martín M, Sabater B. 2010. Plastid ndh genes in plant evolution. *Plant Physiology and Biochemistry* **48(8)**:636–645 DOI 10.1016/j.plaphy.2010.04.009.

Messina R. 2010. `Olea europaea chloroplast, complete genome`. *Available at http://www.ncbi.nlm.nih.gov/nuccore/NC_013707.2* (accessed on 30 June 2015).

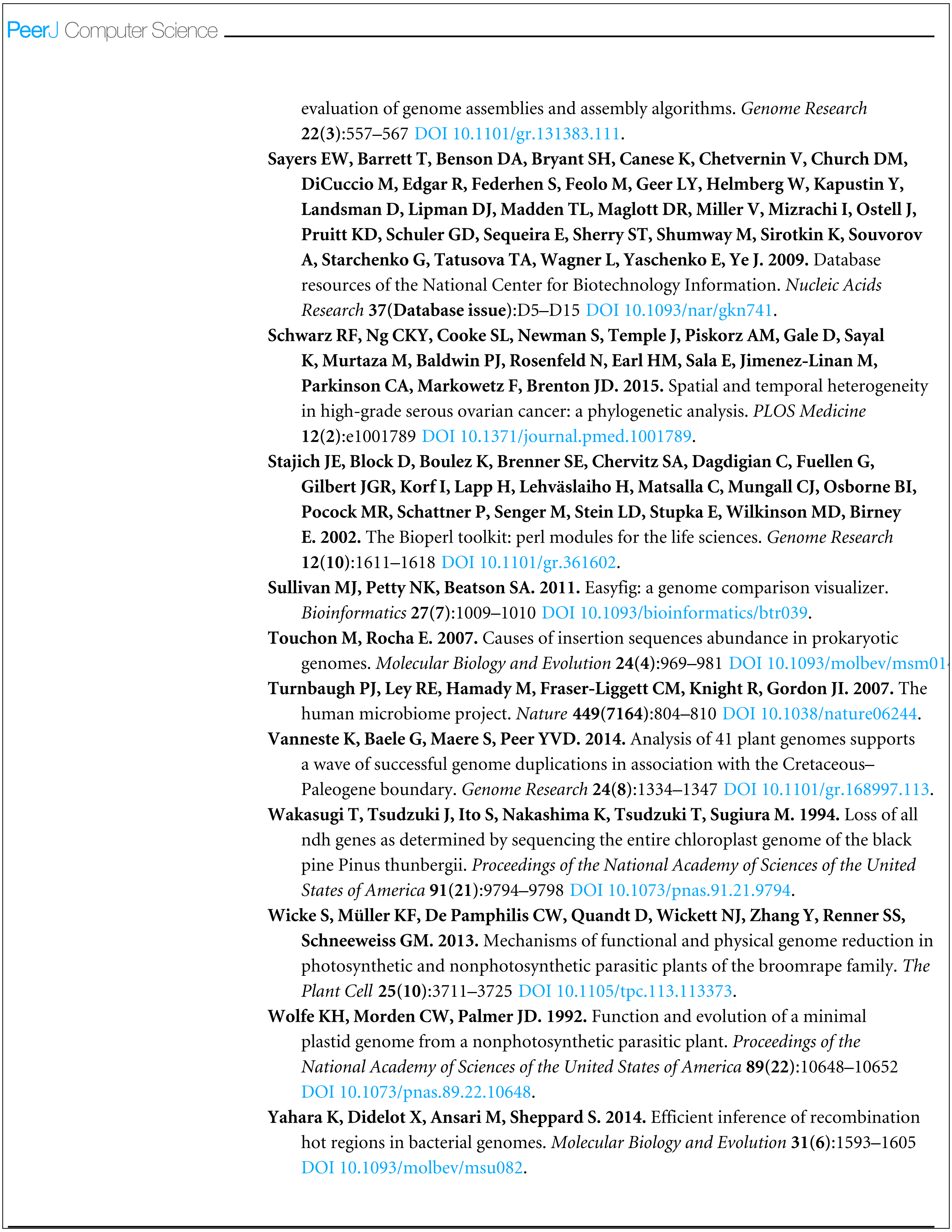
Namouchi A, Didelot X, Schöck U, Gicquel B. 2012. After the bottleneck: genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Research* **22**:721–734 DOI 10.1101/gr.129544.111.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The genomes online database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40(Database issue)**:D571–D579 DOI 10.1093/nar/gkr1100.

Rasko DA, Dale WR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-møller J, Struve C, Petersen AM, Krogfeld KA, Nataro JP, Schadt EE, Waldor MK. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England Journal of Medicine* **365(8)**:709–717 DOI 10.1056/NEJMoa1106920.

Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H. 2012. Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics* **28(2)**:160–166 DOI 10.1093/bioinformatics/btr652.

Salzberg SL, Phillippy AM, Zimin AV, Puiu D, Magoc T, Koren S, Treangen T, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. 2011. GAGE: a critical



evaluation of genome assemblies and assembly algorithms. *Genome Research* **22(3)**:557–567 DOI 10.1101/gr.131383.111.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37(Database issue)**:D5–D15 DOI 10.1093/nar/gkn741.

Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, Gale D, Sayal K, Murtaza M, Baldwin PJ, Rosenfeld N, Earl HM, Sala E, Jimenez-Linan M, Parkinson CA, Markowetz F, Brenton JD. 2015. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLOS Medicine* **12(2)**:e1001789 DOI 10.1371/journal.pmed.1001789.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: perl modules for the life sciences. *Genome Research* **12(10)**:1611–1618 DOI 10.1101/gr.361602.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* **27(7)**:1009–1010 DOI 10.1093/bioinformatics/btr039.

Touchon M, Rocha E. 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution* **24(4)**:969–981 DOI 10.1093/molbev/msm01

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449(7164)**:804–810 DOI 10.1038/nature06244.

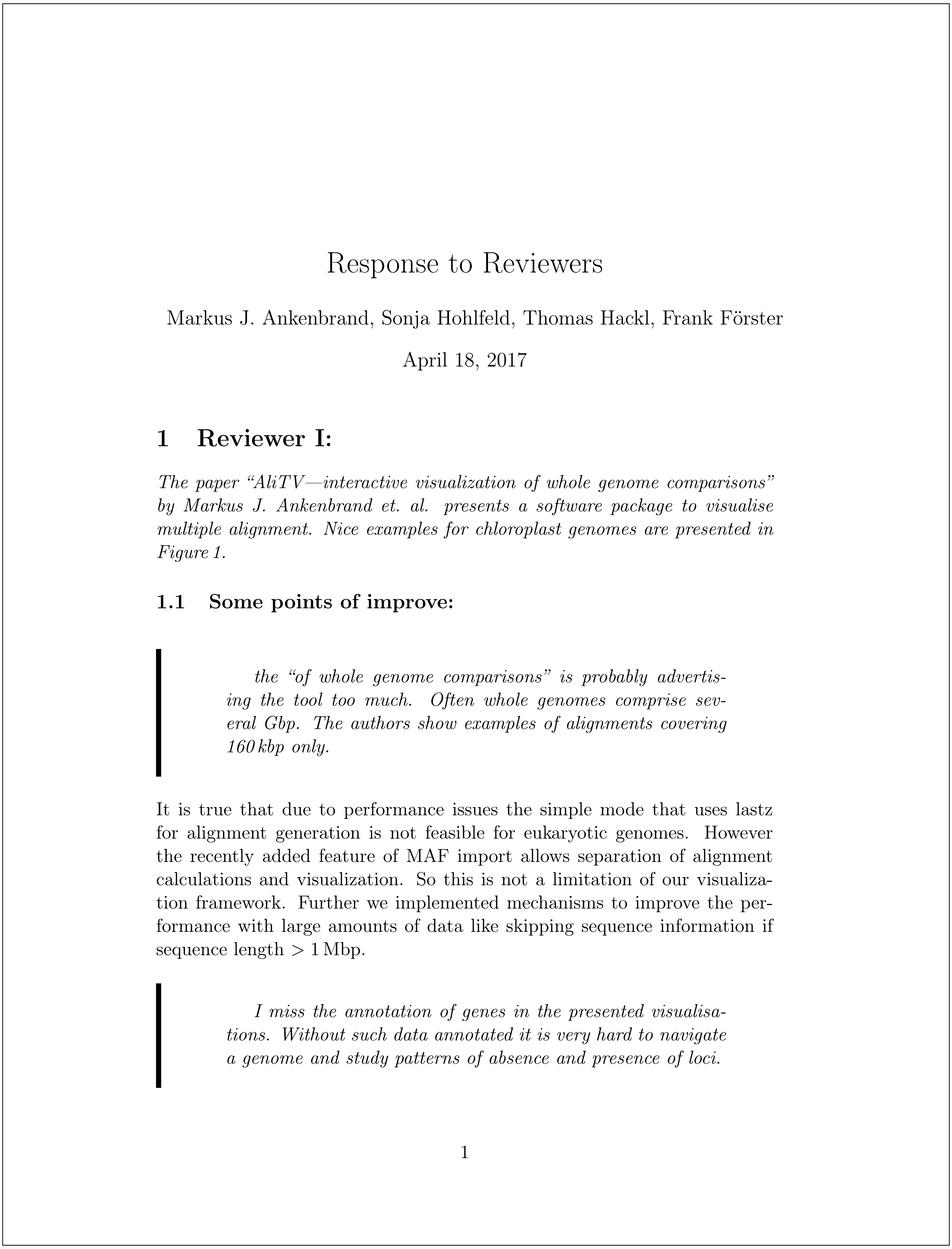
Vanneste K, Baele G, Maere S, Peer YVD. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research* **24(8)**:1334–1347 DOI 10.1101/gr.168997.113.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. *Proceedings of the National Academy of Sciences of the United States of America* **91(21)**:9794–9798 DOI 10.1073/pnas.91.21.9794.

Wicke S, Müller KF, De Pamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *The Plant Cell* **25(10)**:3711–3725 DOI 10.1105/tpc.113.113373.

Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences of the United States of America* **89(22)**:10648–10652 DOI 10.1073/pnas.89.22.10648.

Yahara K, Didelot X, Ansari M, Sheppard S. 2014. Efficient inference of recombination hot regions in bacterial genomes. *Molecular Biology and Evolution* **31(6)**:1593–1605 DOI 10.1093/molbev/msu082.



Response to Reviewers

Markus J. Ankenbrand, Sonja Hohlfeld, Thomas Hackl, Frank Förster

April 18, 2017

1 Reviewer I:

The paper "AliTV—interactive visualization of whole genome comparisons" by Markus J. Ankenbrand et. al. presents a software package to visualise multiple alignment. Nice examples for chloroplast genomes are presented in Figure 1.

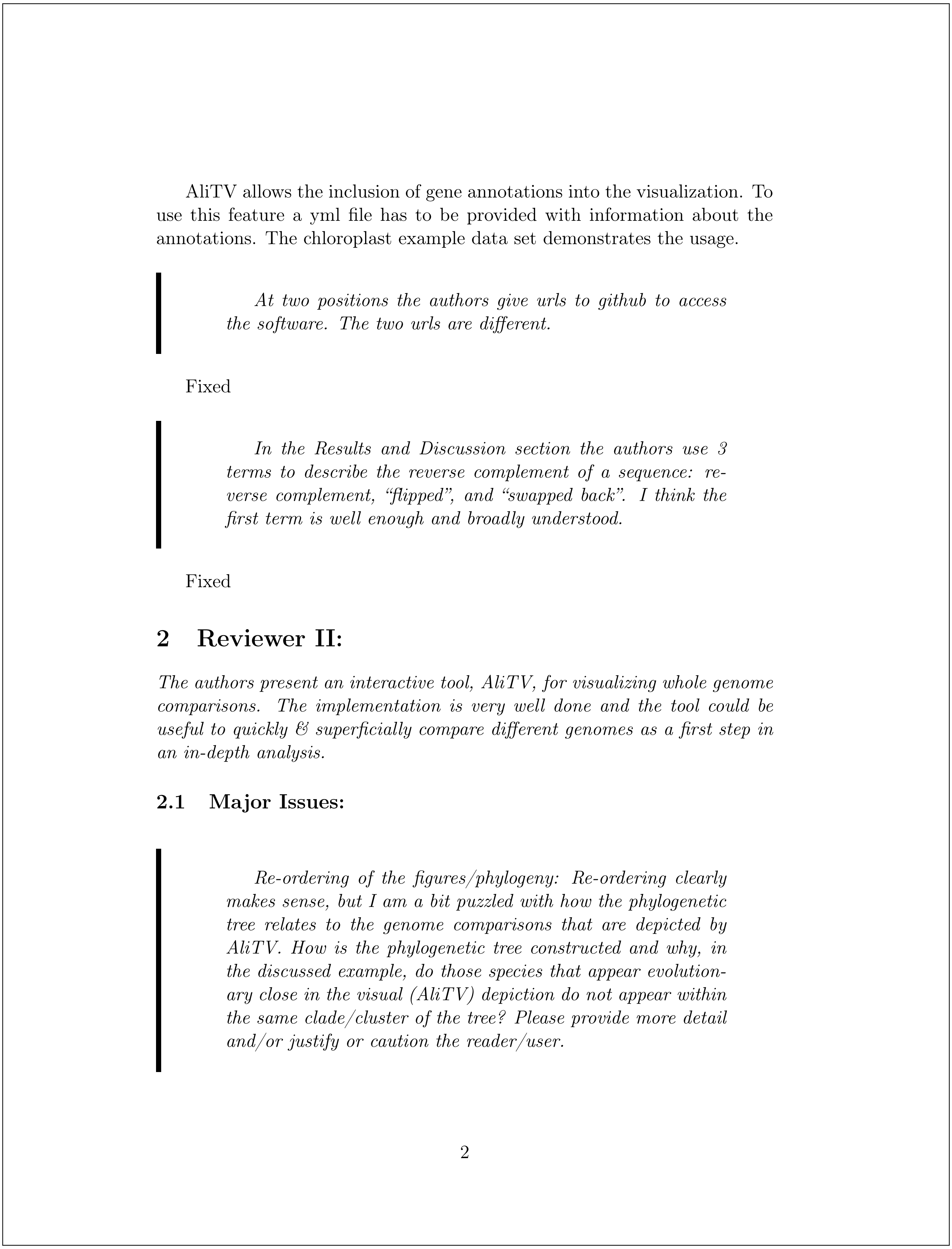
1.1 Some points of improve:

> *the "of whole genome comparisons" is probably advertising the tool too much. Often whole genomes comprise several Gbp. The authors show examples of alignments covering 160 kbp only.*

It is true that due to performance issues the simple mode that uses lastz for alignment generation is not feasible for eukaryotic genomes. However the recently added feature of MAF import allows separation of alignment calculations and visualization. So this is not a limitation of our visualization framework. Further we implemented mechanisms to improve the performance with large amounts of data like skipping sequence information if sequence length > 1 Mbp.

> *I miss the annotation of genes in the presented visualisations. Without such data annotated it is very hard to navigate a genome and study patterns of absence and presence of loci.*

1



AliTV allows the inclusion of gene annotations into the visualization. To use this feature a yml file has to be provided with information about the annotations. The chloroplast example data set demonstrates the usage.

> *At two positions the authors give urls to github to access the software. The two urls are different.*

Fixed

> *In the Results and Discussion section the authors use 3 terms to describe the reverse complement of a sequence: reverse complement, "flipped", and "swapped back". I think the first term is well enough and broadly understood.*

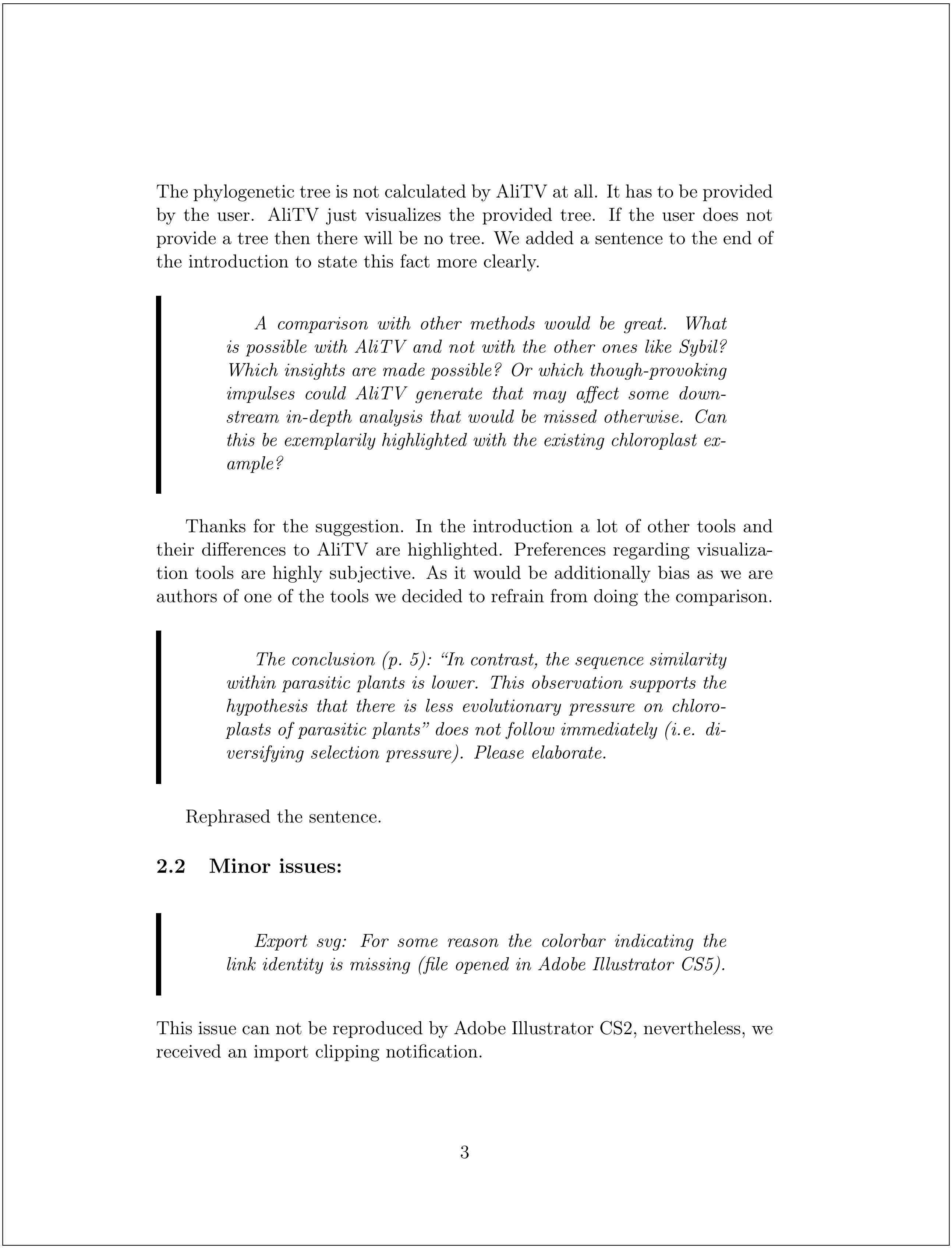
Fixed

2 Reviewer II:

The authors present an interactive tool, AliTV, for visualizing whole genome comparisons. The implementation is very well done and the tool could be useful to quickly & superficially compare different genomes as a first step in an in-depth analysis.

2.1 Major Issues:

> *Re-ordering of the figures/phylogeny: Re-ordering clearly makes sense, but I am a bit puzzled with how the phylogenetic tree relates to the genome comparisons that are depicted by AliTV. How is the phylogenetic tree constructed and why, in the discussed example, do those species that appear evolutionary close in the visual (AliTV) depiction do not appear within the same clade/cluster of the tree? Please provide more detail and/or justify or caution the reader/user.*



The phylogenetic tree is not calculated by AliTV at all. It has to be provided by the user. AliTV just visualizes the provided tree. If the user does not provide a tree then there will be no tree. We added a sentence to the end of the introduction to state this fact more clearly.

> *A comparison with other methods would be great. What is possible with AliTV and not with the other ones like Sybil? Which insights are made possible? Or which though-provoking impulses could AliTV generate that may affect some downstream in-depth analysis that would be missed otherwise. Can this be exemplarily highlighted with the existing chloroplast example?*

Thanks for the suggestion. In the introduction a lot of other tools and their differences to AliTV are highlighted. Preferences regarding visualization tools are highly subjective. As it would be additionally bias as we are authors of one of the tools we decided to refrain from doing the comparison.

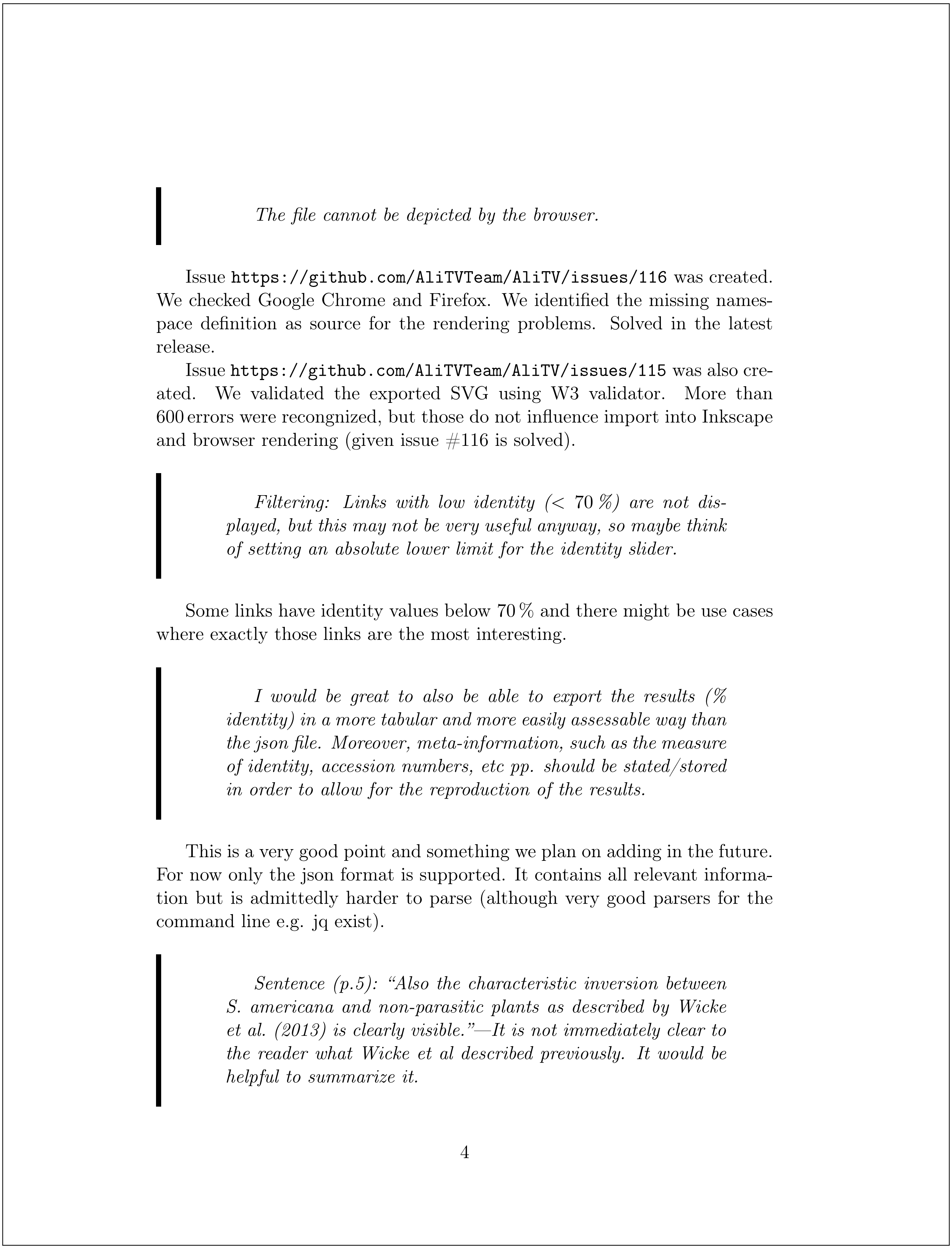
> *The conclusion (p. 5): "In contrast, the sequence similarity within parasitic plants is lower. This observation supports the hypothesis that there is less evolutionary pressure on chloroplasts of parasitic plants" does not follow immediately (i.e. diversifying selection pressure). Please elaborate.*

Rephrased the sentence.

2.2 Minor issues:

> *Export svg: For some reason the colorbar indicating the link identity is missing (file opened in Adobe Illustrator CS5).*

This issue can not be reproduced by Adobe Illustrator CS2, nevertheless, we received an import clipping notification.



> *The file cannot be depicted by the browser.*

Issue `https://github.com/AliTVTeam/AliTV/issues/116` was created. We checked Google Chrome and Firefox. We identified the missing namespace definition as source for the rendering problems. Solved in the latest release.

Issue `https://github.com/AliTVTeam/AliTV/issues/115` was also created. We validated the exported SVG using W3 validator. More than 600 errors were recongnized, but those do not influence import into Inkscape and browser rendering (given issue #116 is solved).

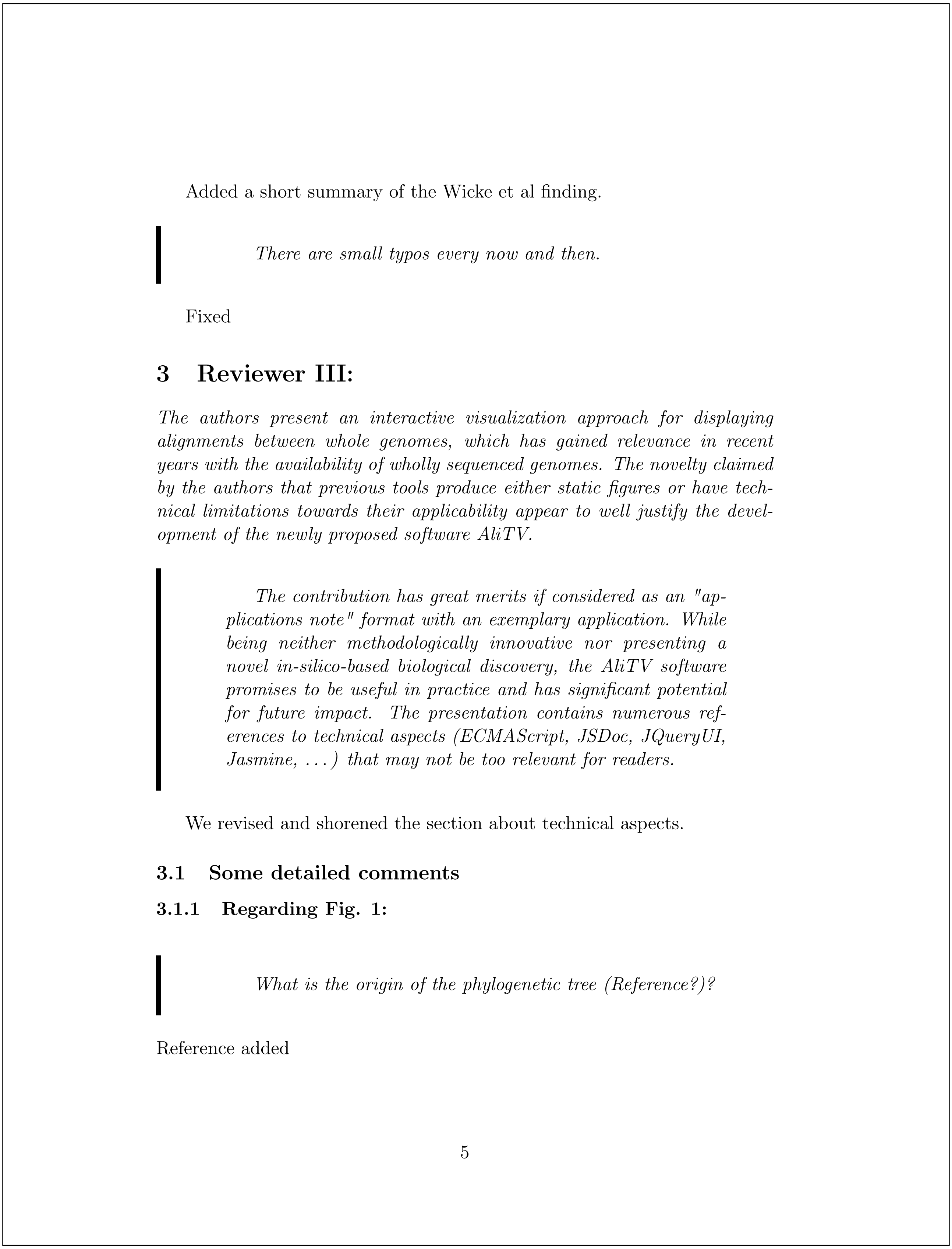
> *Filtering: Links with low identity (< 70 %) are not displayed, but this may not be very useful anyway, so maybe think of setting an absolute lower limit for the identity slider.*

Some links have identity values below 70 % and there might be use cases where exactly those links are the most interesting.

> *I would be great to also be able to export the results (% identity) in a more tabular and more easily assessable way than the json file. Moreover, meta-information, such as the measure of identity, accession numbers, etc pp. should be stated/stored in order to allow for the reproduction of the results.*

This is a very good point and something we plan on adding in the future. For now only the json format is supported. It contains all relevant information but is admittedly harder to parse (although very good parsers for the command line e.g. jq exist).

> *Sentence (p.5): "Also the characteristic inversion between S. americana and non-parasitic plants as described by Wicke et al. (2013) is clearly visible."—It is not immediately clear to the reader what Wicke et al described previously. It would be helpful to summarize it.*



Added a short summary of the Wicke et al finding.

> *There are small typos every now and then.*

Fixed

3 Reviewer III:

The authors present an interactive visualization approach for displaying alignments between whole genomes, which has gained relevance in recent years with the availability of wholly sequenced genomes. The novelty claimed by the authors that previous tools produce either static figures or have technical limitations towards their applicability appear to well justify the development of the newly proposed software AliTV.

> *The contribution has great merits if considered as an "applications note" format with an exemplary application. While being neither methodologically innovative nor presenting a novel in-silico-based biological discovery, the AliTV software promises to be useful in practice and has significant potential for future impact. The presentation contains numerous references to technical aspects (ECMAScript, JSDoc, JQueryUI, Jasmine, . . .) that may not be too relevant for readers.*

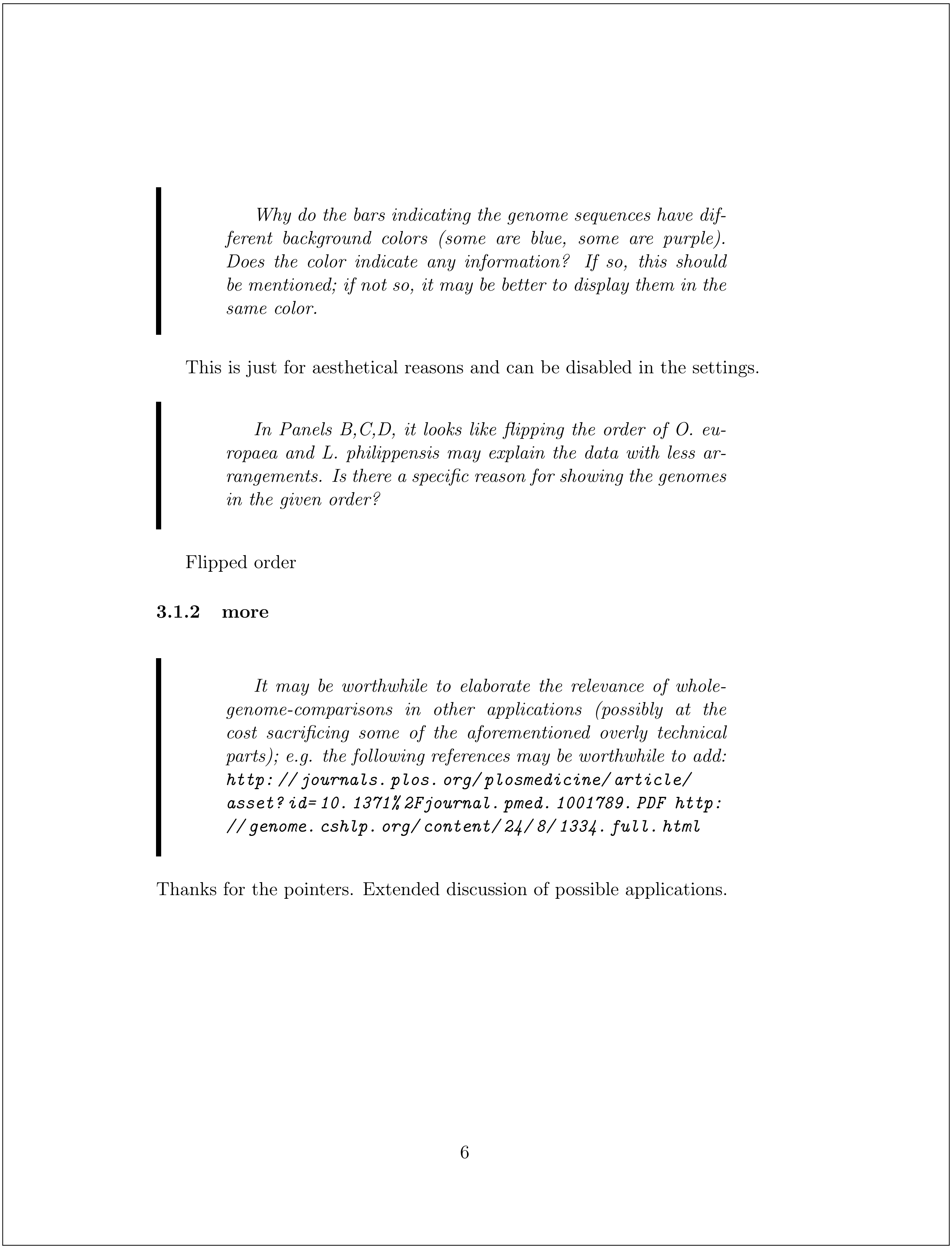
We revised and shorened the section about technical aspects.

3.1 Some detailed comments

3.1.1 Regarding Fig. 1:

> *What is the origin of the phylogenetic tree (Reference?)?*

Reference added



> *Why do the bars indicating the genome sequences have different background colors (some are blue, some are purple). Does the color indicate any information? If so, this should be mentioned; if not so, it may be better to display them in the same color.*

This is just for aesthetical reasons and can be disabled in the settings.

> *In Panels B,C,D, it looks like flipping the order of O. europaea and L. philippensis may explain the data with less arrangements. Is there a specific reason for showing the genomes in the given order?*

Flipped order

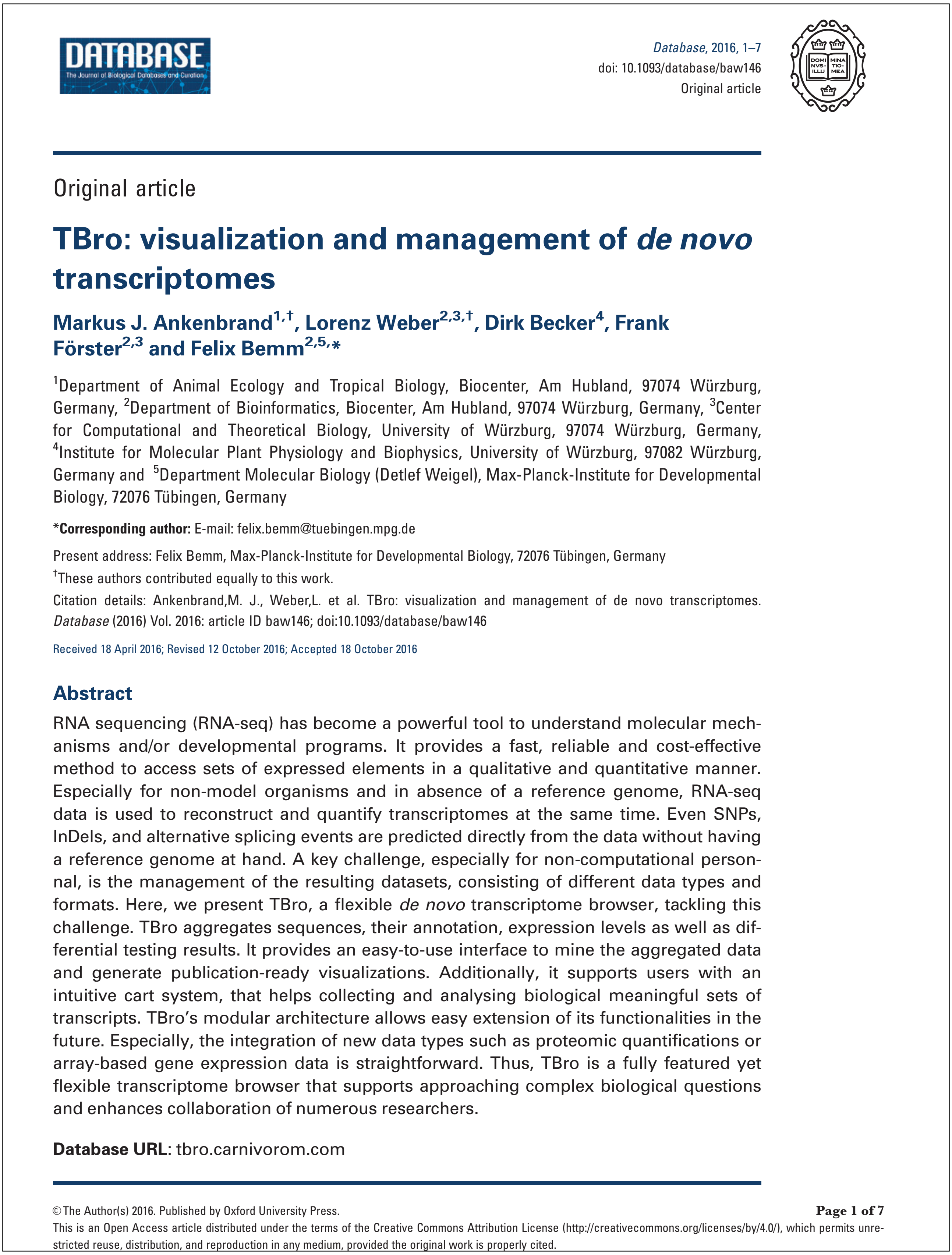
3.1.2 more

> *It may be worthwhile to elaborate the relevance of whole-genome-comparisons in other applications (possibly at the cost sacrificing some of the aforementioned overly technical parts); e.g. the following references may be worthwhile to add:* *http://journals.plos.org/plosmedicine/article/asset?id=10.1371%2Fjournal.pmed.1001789.PDF http://genome.cshlp.org/content/24/8/1334.full.html*

Thanks for the pointers. Extended discussion of possible applications.

5.11 TBRO: VISUALIZATION AND MANAGEMENT OF DE NOVO TRANSCRIPTOMES

– published in *Database* –



Database, 2016, 1–7
doi: 10.1093/database/baw146
Original article

Original article

TBro: visualization and management of *de novo* transcriptomes

Markus J. Ankenbrand[1,†], Lorenz Weber[2,3,†], Dirk Becker[4], Frank Förster[2,3] and Felix Bemm[2,5,*]

[1]Department of Animal Ecology and Tropical Biology, Biocenter, Am Hubland, 97074 Würzburg, Germany, [2]Department of Bioinformatics, Biocenter, Am Hubland, 97074 Würzburg, Germany, [3]Center for Computational and Theoretical Biology, University of Würzburg, 97074 Würzburg, Germany, [4]Institute for Molecular Plant Physiology and Biophysics, University of Würzburg, 97082 Würzburg, Germany and [5]Department Molecular Biology (Detlef Weigel), Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

***Corresponding author:** E-mail: felix.bemm@tuebingen.mpg.de

Present address: Felix Bemm, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

[†]These authors contributed equally to this work.

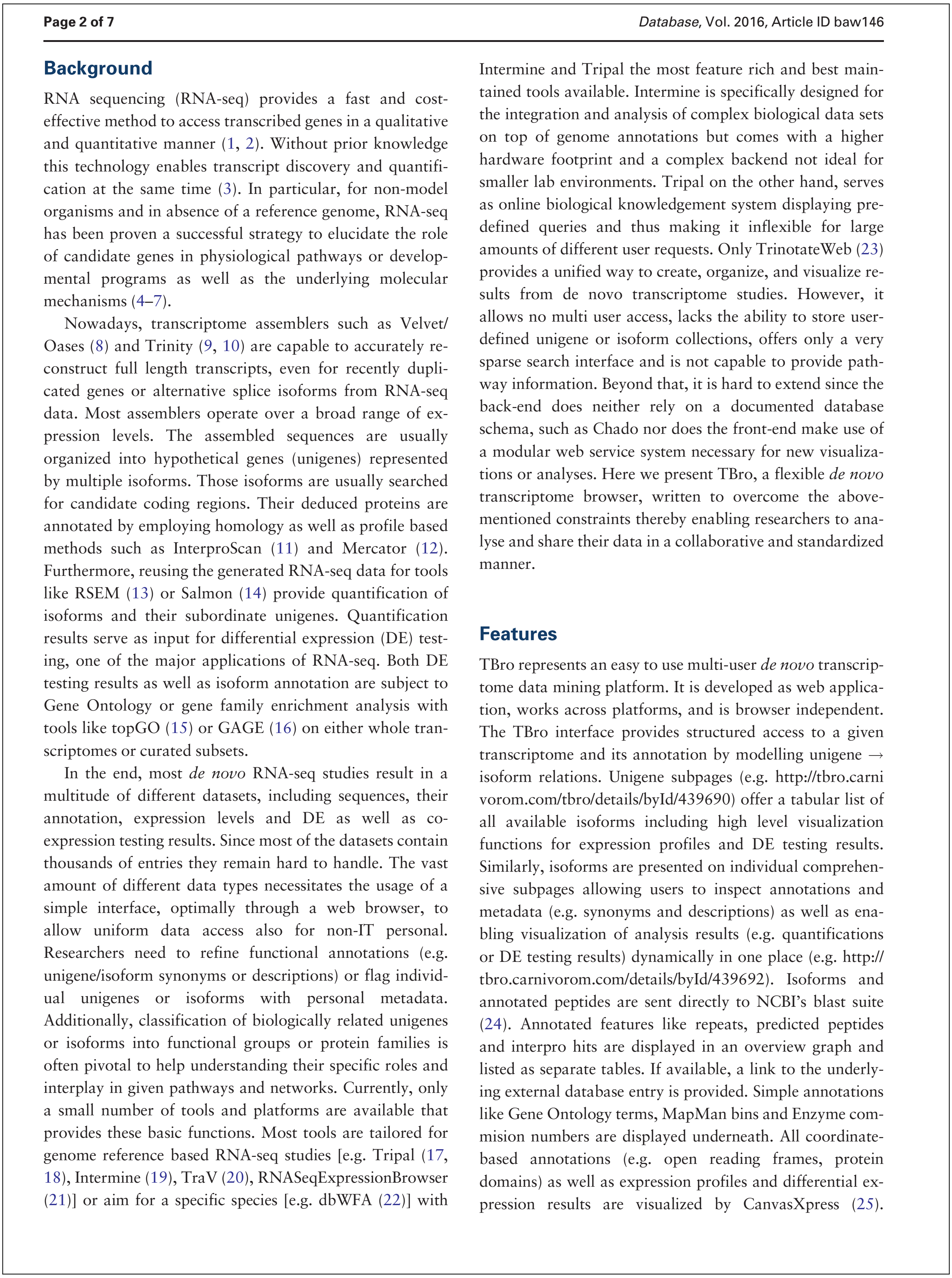
Citation details: Ankenbrand,M. J., Weber,L. et al. TBro: visualization and management of de novo transcriptomes. *Database* (2016) Vol. 2016: article ID baw146; doi:10.1093/database/baw146

Received 18 April 2016; Revised 12 October 2016; Accepted 18 October 2016

Abstract

RNA sequencing (RNA-seq) has become a powerful tool to understand molecular mechanisms and/or developmental programs. It provides a fast, reliable and cost-effective method to access sets of expressed elements in a qualitative and quantitative manner. Especially for non-model organisms and in absence of a reference genome, RNA-seq data is used to reconstruct and quantify transcriptomes at the same time. Even SNPs, InDels, and alternative splicing events are predicted directly from the data without having a reference genome at hand. A key challenge, especially for non-computational personnal, is the management of the resulting datasets, consisting of different data types and formats. Here, we present TBro, a flexible *de novo* transcriptome browser, tackling this challenge. TBro aggregates sequences, their annotation, expression levels as well as differential testing results. It provides an easy-to-use interface to mine the aggregated data and generate publication-ready visualizations. Additionally, it supports users with an intuitive cart system, that helps collecting and analysing biological meaningful sets of transcripts. TBro's modular architecture allows easy extension of its functionalities in the future. Especially, the integration of new data types such as proteomic quantifications or array-based gene expression data is straightforward. Thus, TBro is a fully featured yet flexible transcriptome browser that supports approaching complex biological questions and enhances collaboration of numerous researchers.

Database URL: tbro.carnivorom.com



Background

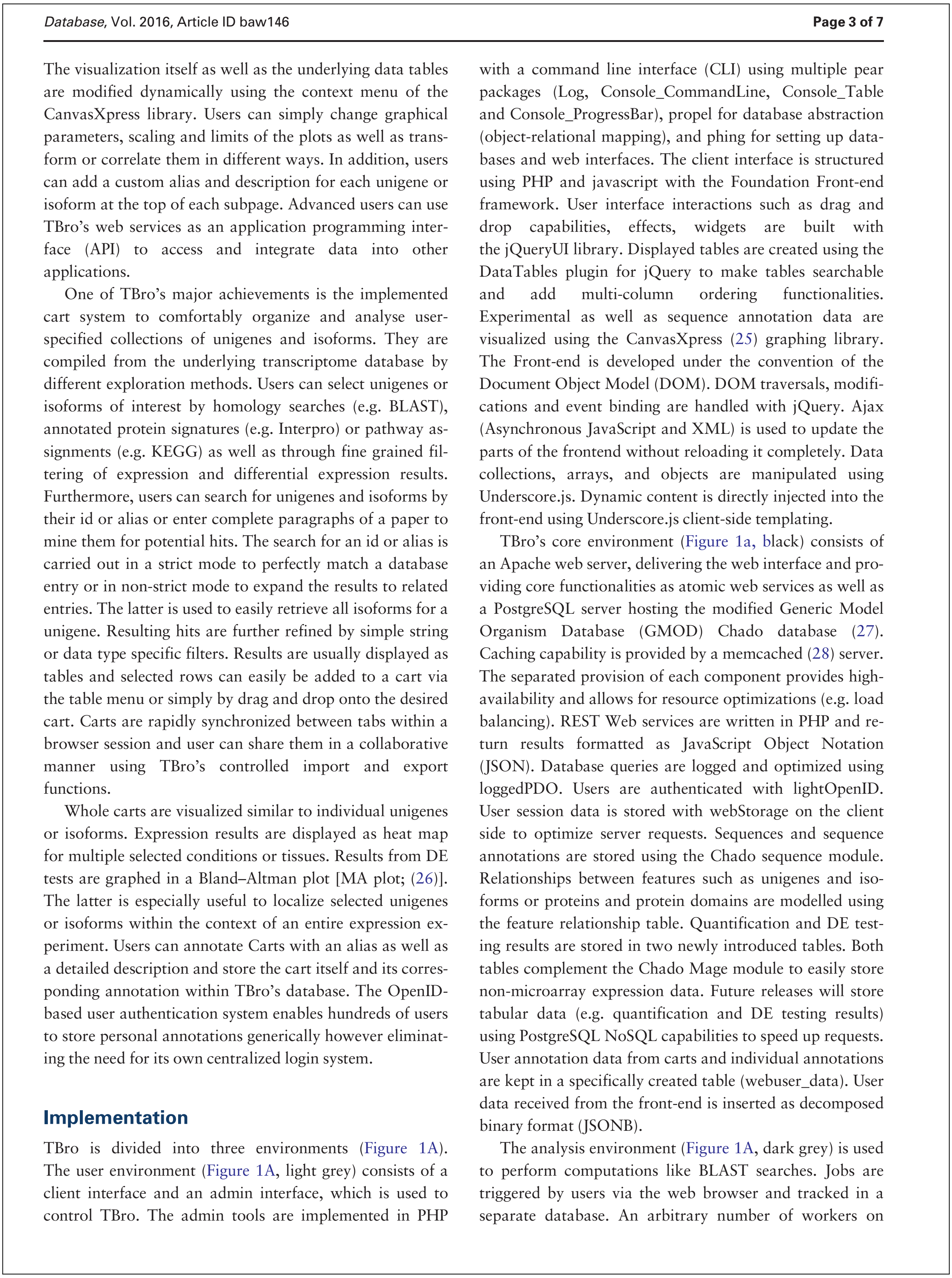
RNA sequencing (RNA-seq) provides a fast and cost-effective method to access transcribed genes in a qualitative and quantitative manner (1, 2). Without prior knowledge this technology enables transcript discovery and quantification at the same time (3). In particular, for non-model organisms and in absence of a reference genome, RNA-seq has been proven a successful strategy to elucidate the role of candidate genes in physiological pathways or developmental programs as well as the underlying molecular mechanisms (4–7).

Nowadays, transcriptome assemblers such as Velvet/Oases (8) and Trinity (9, 10) are capable to accurately reconstruct full length transcripts, even for recently duplicated genes or alternative splice isoforms from RNA-seq data. Most assemblers operate over a broad range of expression levels. The assembled sequences are usually organized into hypothetical genes (unigenes) represented by multiple isoforms. Those isoforms are usually searched for candidate coding regions. Their deduced proteins are annotated by employing homology as well as profile based methods such as InterproScan (11) and Mercator (12). Furthermore, reusing the generated RNA-seq data for tools like RSEM (13) or Salmon (14) provide quantification of isoforms and their subordinate unigenes. Quantification results serve as input for differential expression (DE) testing, one of the major applications of RNA-seq. Both DE testing results as well as isoform annotation are subject to Gene Ontology or gene family enrichment analysis with tools like topGO (15) or GAGE (16) on either whole transcriptomes or curated subsets.

In the end, most *de novo* RNA-seq studies result in a multitude of different datasets, including sequences, their annotation, expression levels and DE as well as co-expression testing results. Since most of the datasets contain thousands of entries they remain hard to handle. The vast amount of different data types necessitates the usage of a simple interface, optimally through a web browser, to allow uniform data access also for non-IT personal. Researchers need to refine functional annotations (e.g. unigene/isoform synonyms or descriptions) or flag individual unigenes or isoforms with personal metadata. Additionally, classification of biologically related unigenes or isoforms into functional groups or protein families is often pivotal to help understanding their specific roles and interplay in given pathways and networks. Currently, only a small number of tools and platforms are available that provides these basic functions. Most tools are tailored for genome reference based RNA-seq studies [e.g. Tripal (17, 18), Intermine (19), TraV (20), RNASeqExpressionBrowser (21)] or aim for a specific species [e.g. dbWFA (22)] with Intermine and Tripal the most feature rich and best maintained tools available. Intermine is specifically designed for the integration and analysis of complex biological data sets on top of genome annotations but comes with a higher hardware footprint and a complex backend not ideal for smaller lab environments. Tripal on the other hand, serves as online biological knowledgement system displaying predefined queries and thus making it inflexible for large amounts of different user requests. Only TrinotateWeb (23) provides a unified way to create, organize, and visualize results from de novo transcriptome studies. However, it allows no multi user access, lacks the ability to store user-defined unigene or isoform collections, offers only a very sparse search interface and is not capable to provide pathway information. Beyond that, it is hard to extend since the back-end does neither rely on a documented database schema, such as Chado nor does the front-end make use of a modular web service system necessary for new visualizations or analyses. Here we present TBro, a flexible *de novo* transcriptome browser, written to overcome the above-mentioned constraints thereby enabling researchers to analyse and share their data in a collaborative and standardized manner.

Features

TBro represents an easy to use multi-user *de novo* transcriptome data mining platform. It is developed as web application, works across platforms, and is browser independent. The TBro interface provides structured access to a given transcriptome and its annotation by modelling unigene → isoform relations. Unigene subpages (e.g. http://tbro.carnivorom.com/tbro/details/byId/439690) offer a tabular list of all available isoforms including high level visualization functions for expression profiles and DE testing results. Similarly, isoforms are presented on individual comprehensive subpages allowing users to inspect annotations and metadata (e.g. synonyms and descriptions) as well as enabling visualization of analysis results (e.g. quantifications or DE testing results) dynamically in one place (e.g. http://tbro.carnivorom.com/details/byId/439692). Isoforms and annotated peptides are sent directly to NCBI's blast suite (24). Annotated features like repeats, predicted peptides and interpro hits are displayed in an overview graph and listed as separate tables. If available, a link to the underlying external database entry is provided. Simple annotations like Gene Ontology terms, MapMan bins and Enzyme commision numbers are displayed underneath. All coordinate-based annotations (e.g. open reading frames, protein domains) as well as expression profiles and differential expression results are visualized by CanvasXpress (25).



The visualization itself as well as the underlying data tables are modified dynamically using the context menu of the CanvasXpress library. Users can simply change graphical parameters, scaling and limits of the plots as well as transform or correlate them in different ways. In addition, users can add a custom alias and description for each unigene or isoform at the top of each subpage. Advanced users can use TBro's web services as an application programming interface (API) to access and integrate data into other applications.

One of TBro's major achievements is the implemented cart system to comfortably organize and analyse user-specified collections of unigenes and isoforms. They are compiled from the underlying transcriptome database by different exploration methods. Users can select unigenes or isoforms of interest by homology searches (e.g. BLAST), annotated protein signatures (e.g. Interpro) or pathway assignments (e.g. KEGG) as well as through fine grained filtering of expression and differential expression results. Furthermore, users can search for unigenes and isoforms by their id or alias or enter complete paragraphs of a paper to mine them for potential hits. The search for an id or alias is carried out in a strict mode to perfectly match a database entry or in non-strict mode to expand the results to related entries. The latter is used to easily retrieve all isoforms for a unigene. Resulting hits are further refined by simple string or data type specific filters. Results are usually displayed as tables and selected rows can easily be added to a cart via the table menu or simply by drag and drop onto the desired cart. Carts are rapidly synchronized between tabs within a browser session and user can share them in a collaborative manner using TBro's controlled import and export functions.

Whole carts are visualized similar to individual unigenes or isoforms. Expression results are displayed as heat map for multiple selected conditions or tissues. Results from DE tests are graphed in a Bland–Altman plot [MA plot; (26)]. The latter is especially useful to localize selected unigenes or isoforms within the context of an entire expression experiment. Users can annotate Carts with an alias as well as a detailed description and store the cart itself and its corresponding annotation within TBro's database. The OpenID-based user authentication system enables hundreds of users to store personal annotations generically however eliminating the need for its own centralized login system.

Implementation

TBro is divided into three environments (Figure 1A). The user environment (Figure 1A, light grey) consists of a client interface and an admin interface, which is used to control TBro. The admin tools are implemented in PHP with a command line interface (CLI) using multiple pear packages (Log, Console_CommandLine, Console_Table and Console_ProgressBar), propel for database abstraction (object-relational mapping), and phing for setting up databases and web interfaces. The client interface is structured using PHP and javascript with the Foundation Front-end framework. User interface interactions such as drag and drop capabilities, effects, widgets are built with the jQueryUI library. Displayed tables are created using the DataTables plugin for jQuery to make tables searchable and add multi-column ordering functionalities. Experimental as well as sequence annotation data are visualized using the CanvasXpress (25) graphing library. The Front-end is developed under the convention of the Document Object Model (DOM). DOM traversals, modifications and event binding are handled with jQuery. Ajax (Asynchronous JavaScript and XML) is used to update the parts of the frontend without reloading it completely. Data collections, arrays, and objects are manipulated using Underscore.js. Dynamic content is directly injected into the front-end using Underscore.js client-side templating.

TBro's core environment (Figure 1a, black) consists of an Apache web server, delivering the web interface and providing core functionalities as atomic web services as well as a PostgreSQL server hosting the modified Generic Model Organism Database (GMOD) Chado database (27). Caching capability is provided by a memcached (28) server. The separated provision of each component provides high-availability and allows for resource optimizations (e.g. load balancing). REST Web services are written in PHP and return results formatted as JavaScript Object Notation (JSON). Database queries are logged and optimized using loggedPDO. Users are authenticated with lightOpenID. User session data is stored with webStorage on the client side to optimize server requests. Sequences and sequence annotations are stored using the Chado sequence module. Relationships between features such as unigenes and isoforms or proteins and protein domains are modelled using the feature relationship table. Quantification and DE testing results are stored in two newly introduced tables. Both tables complement the Chado Mage module to easily store non-microarray expression data. Future releases will store tabular data (e.g. quantification and DE testing results) using PostgreSQL NoSQL capabilities to speed up requests. User annotation data from carts and individual annotations are kept in a specifically created table (webuser_data). User data received from the front-end is inserted as decomposed binary format (JSONB).

The analysis environment (Figure 1A, dark grey) is used to perform computations like BLAST searches. Jobs are triggered by users via the web browser and tracked in a separate database. An arbitrary number of workers on

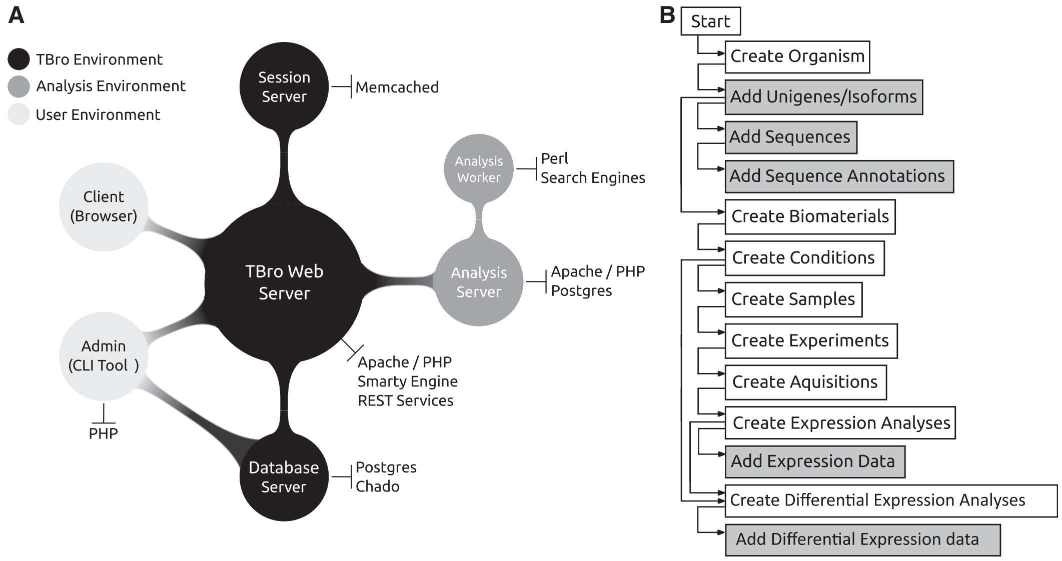


Figure 1. (A) TBro's architecture is divided into three sections. The TBro environment builds the backbone with the central web server. The web server is connected to the database server and the session server for caching. The analysis environment is used to perform computationally intensive tasks. It is divided into a server and an arbitrary number of workers that can run on heterogeneous systems. The user environment consists of the client (a web browser) which is used to interact with a running instance of TBro and the command line tools which are used to import and manage data by a qualified administrator. (B) A typical data import hierarchically prepares and adds all transcriptomic data sets. Tasks performed by TBro-db are coloured in grey while tasks performed with TBro-import are coloured in white. The complete workflow tightly builds on the reference Chado schema to ease maintenance and usability.

heterogeneous host systems (currently Linux and Windows are supported) is utilized to run the job. Workers query the database for unallocated jobs, run them and report the results back to the database. The status of the job and eventually the results are accessible by the user via a unique URL. The analysis environment builds on a modular structure to easily extend it to other tools (e.g. HMMER for profile based searches).

Usage

TBro knows two principal roles: administrator and user. The administrator imports and manages data using a CLI while the user accesses and searches the data with a web browser. The CLI is divided into three subcommands, TBro-db for managing data values (list, insert, edit and delete of e.g. contacts, organisms), TBro-import for importing multiple data values from files (e.g. ids, sequences) and TBro-tool which provides helper scripts (e.g. format converter). All tools come with support for auto completion in Linux environments. The CLI tools hierarchically prepare and import all data sets but can also be used to retrieve data from the database. An exemplary import workflow is available in the TBro documentation (http://tbro-tutorial.readthedocs.org). Sequence information and relations are imported by supplying relation maps (Unigene → Isoforms and Isoform → Open Reading Frame) and simple fasta files. The same is done for generic pathway associations (EC → KEGG Map). Annotation results are imported using a two-column tab-separated file (Sequence ID → GO/EC/Synonym) or source-defined multi-column files (Interpro, RepeatMasker, MapMan). Expression counts and DE results are imported after deep modelling the sample relations with TBro's database control tool (TBro-db, Figure 2B). Each expression dataset is associated with a biomaterial (e.g. tissue), a condition (e.g. treatment) and a sample name (e.g. replicate-1) according to the Chado database schema. The combination of biomaterial, condition and sample name is connected with an experiment. Each experiment is assigned to one or multiple acquisitions corresponding to a sequencing runs or array hybridization. Acquisitions are associated with a corresponding analysis e.g. quantification and normalization of unigene and isoform counts or DE test results. Finally, the datasets are imported by simply supplying a quantification and analysis id.

The online demo (http://tbro.carnivorom.com) hosts data from the recently published Venus flytrap *(Dionaea muscipula)* deep transcriptome sequencing project (Bemm et al., 2016, in press). The unfiltered data sets contain 315 584 isoforms for 183 578 subordinate unigenes. A total of 3 221 001 annotation entries of various types are stored within TBro's database backend. Expression data

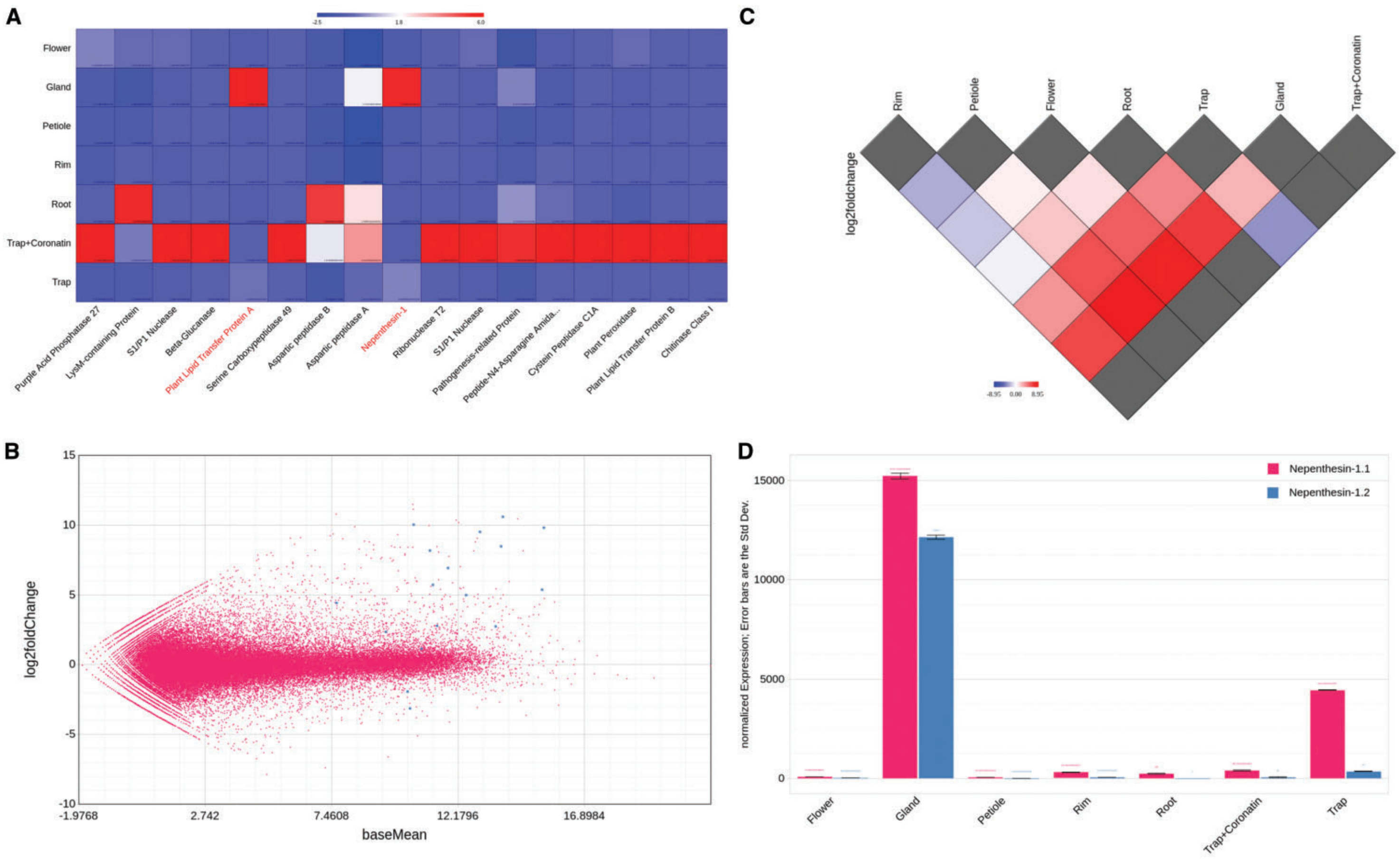


Figure 2. (A) Z-transformed expression heatmap of a cart containing putative members of the hydrolytic cocktails secreted by Venus flytrap during its hunting cycle. Two unigenes are being expressed in a non-stimulated gland specific manner. (B) MA plot for the same cart based on DE testing results from DESeq. The plot indicates that most members of the hydrolytic cocktail are being highly expressed compared to the majority of the unigenes. (C) Triangular visualization of the DE testing results for an individual gene (Nepenthesis-1). (D) Simple expression barplot of the Nepenthesin-1 gene with two isoforms showing different expression patterns. All plots were generated directly in TBro. Z-transformation, scaling and layouts were adjusted using functions from CanvasXpress context menu directly in the browser.

from four experiments with a total of 39 samples contain 19 467 318 distinct expression values and results for 2 744 423 DE comparisons are aggregated. The total size of the PostgreSQL database on disk is approximately 14 GB. All components of the Venus Flytrap TBro instance are running on a single virtual machine [Intel(R) Xeon(R) CPU E5-2640 v3, 2 cores, 8 GB RAM, Ubuntu 12.04, 64 bit].

One of the major questions during the deep transcriptome sequencing project of the Venus flytrap was about the nature and abundance of the hydrolytic enzymes which are secreted by specialized glands on the inner trap surface to digest animal prey. Several high-throughput proteomics experiments using different stimuli (insect and hormone treatment as well as mechanical stimulation) were conducted to stimulate secretion and detect hydrolytic enzymes in *Dionaea's* digestive fluid. Following sampling of the secretion fluid, peptides were identified by mass spectrometry and mapped onto the reference transcriptome. Thereby 368 isoforms, respectively their deduced proteins, were identified as secreted independent of the nature of the stimulus. The resulting isoforms were searched within TBro and stored using its cart system. This initial 'secretome' cart was searched for entries exhibiting an annotated signal peptide (indicative for secreted proteins) employing the cart annotation search. Eligible isoforms were added to a 'filtered secretome' cart. Subordinate unigenes were added to the new 'filtered secretome' cart via the table menu and DE results from insect-stimulated glands (exp008) were visualized using a MA plot (Figure 2B). It became immediately obvious that the hydrolytic cocktail consists of enzymes being already expressed in non-stimulated glands (Figure 2B, blue dots with $\log_2$ fold change < 0, 2 unigenes) and those triggered upon insect stimulation (Figure 2B, blue dots with $\log_2$ fold change > 0, 15 unigenes). The two differentially expressed unigenes in non-stimulated tissues were further analysed with TBro's triangular DE plot using an expression experiment comprising different non-stimulated tissues (exp001, Figure 2C). This plot revealed that the two unigenes, encoding Nepenthesin-1 and a Lipid Transfer Protein (LTP), are indeed excessively transcribed in a gland specific manner. The refined cart was directly used as supplementary data for the publication and to ease the review process.

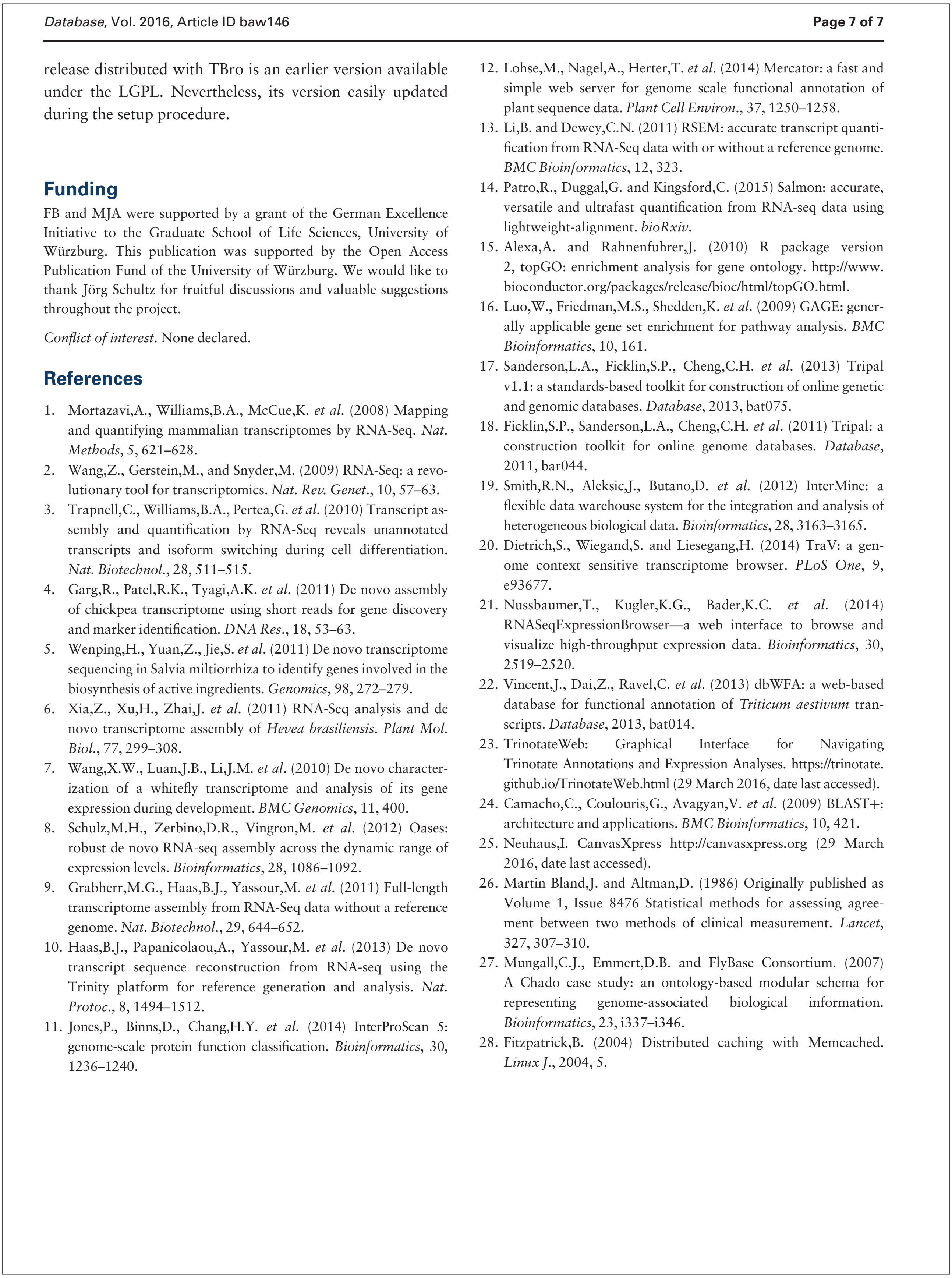
Altogether, TBro successfully enhanced collaboration of numerous researchers working in the Venus flytrap transcriptome project team. It was particularly helpful to visualize expression strength or expression variability in publication-associated carts (Figure 2A). It was further intensively used to identify representative isoforms for individual unigenes using its adjustable expression bar plots (Figure 2D). Researchers frequently visualized DE test results using TBro's triangular DE plot (Figure 2C) to identify DE patterns over a large set of different tissues. Finally, TBro's pathway module was used to provide functional associations (e.g. Jasmonic acid biosynthesis, Supplementary Figure S2).

Conclusion

TBro provides simple-to-use interfaces to (i) inspect and refine functional annotations, (ii) analyse and visualize expression as well as (iii) DE testing data. It handles user derived sets of unigenes/isoforms as well as entire experimental datasets and thus outperforms competing packages in terms of functionality, user-friendliness and flexibility. The cart system helps collecting, organizing and sharing biological meaningful sets of unigenes/isoforms and thus offers an effective way to export meta-data for external review. Building on the Chado database schema empowers TBro to handle complex representations of biological knowledge and a multitude of different data types. Although TBro was developed with RNA-seq experiments in mind, it can easily be adopted to host proteomic or other quantification data sets. Furthermore, it provides interoperability between different biological databases and applications of the GMOD toolkit. The modular backend, organized into different environments and the heavy use of highly flexible atomic services allow an easy extension of TBro's functionalities in the future. It also provides a fast prototyping platform to test and develop functionalities for genome-centred data warehouse systems such as Intermine and Tripal. Upcoming releases will introduce cart operations such as union or intersection as well as transformations (e.g. unigene ↔ isoform) to further ease TBro's usage. Finally, we aim to develop new features that enable users to switch between organisms or data releases in context of their personal carts again using Chado's built-in relationship model.

Availability

TBro is available as docker images (https://hub.docker.com/u/tbroteam) as well as source code (https://github.com/tbroteam). It is easily set-up using preconfigured docker images. Core applications, databases and job handlers are distributed in separate images. Functional tests are continuously performed with Travis-CI (https://travis-ci.org/TBroTeam/TBro) while code review is automatically performed by codeclimate (https://codeclimate.com/github/TBroTeam/TBro). A tutorial leads user through the installation as well as analysis process (https://tbro-tutorial.readthedocs.org). TBro is distributed under the MIT license. All included modules have compatible licenses (see Supplementary Table S1). The CanvasXpress (http://canvasxpress.org)



release distributed with TBro is an earlier version available under the LGPL. Nevertheless, its version easily updated during the setup procedure.

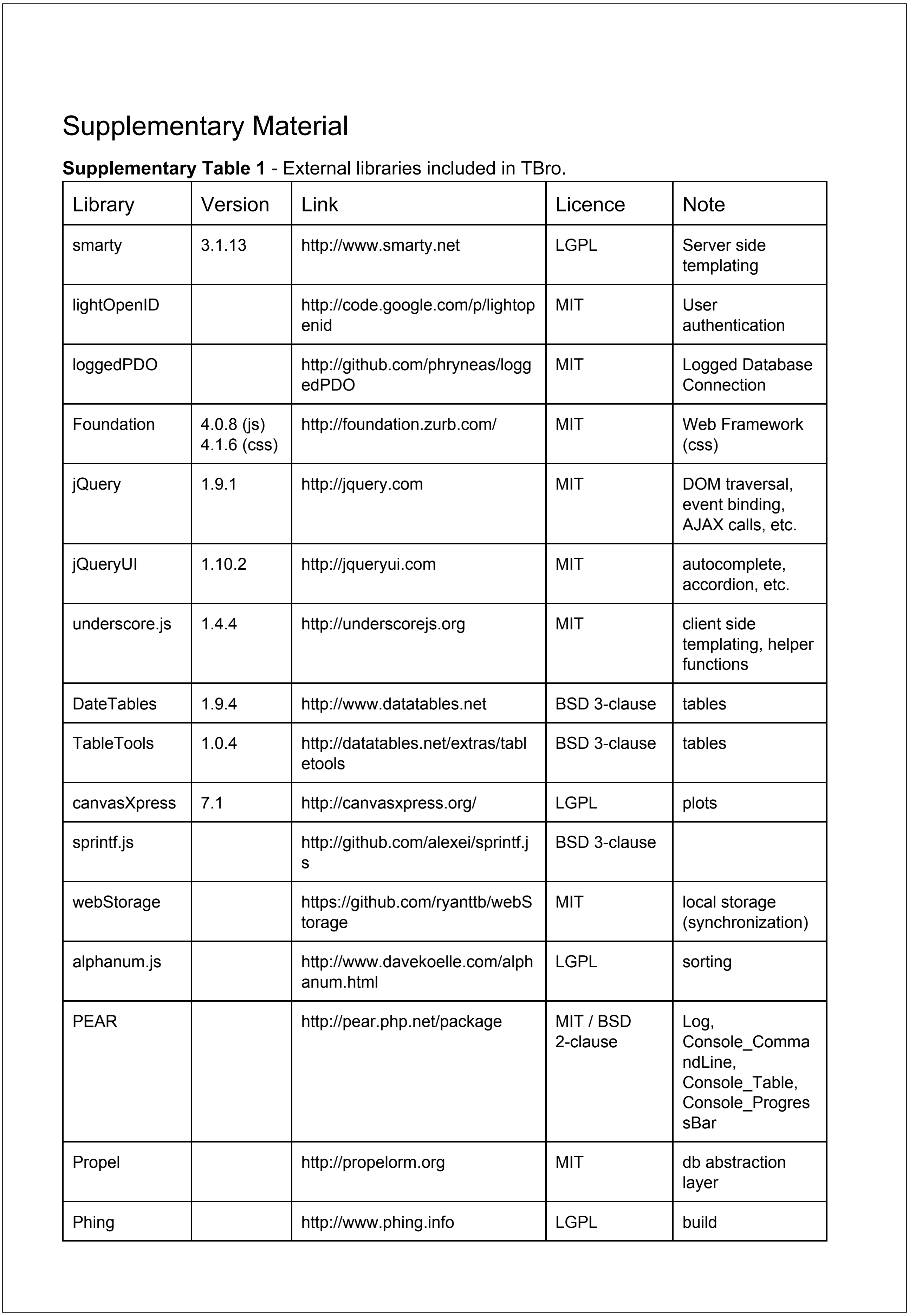
Funding

FB and MJA were supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. This publication was supported by the Open Access Publication Fund of the University of Würzburg. We would like to thank Jörg Schultz for fruitful discussions and valuable suggestions throughout the project.

Conflict of interest. None declared.

References

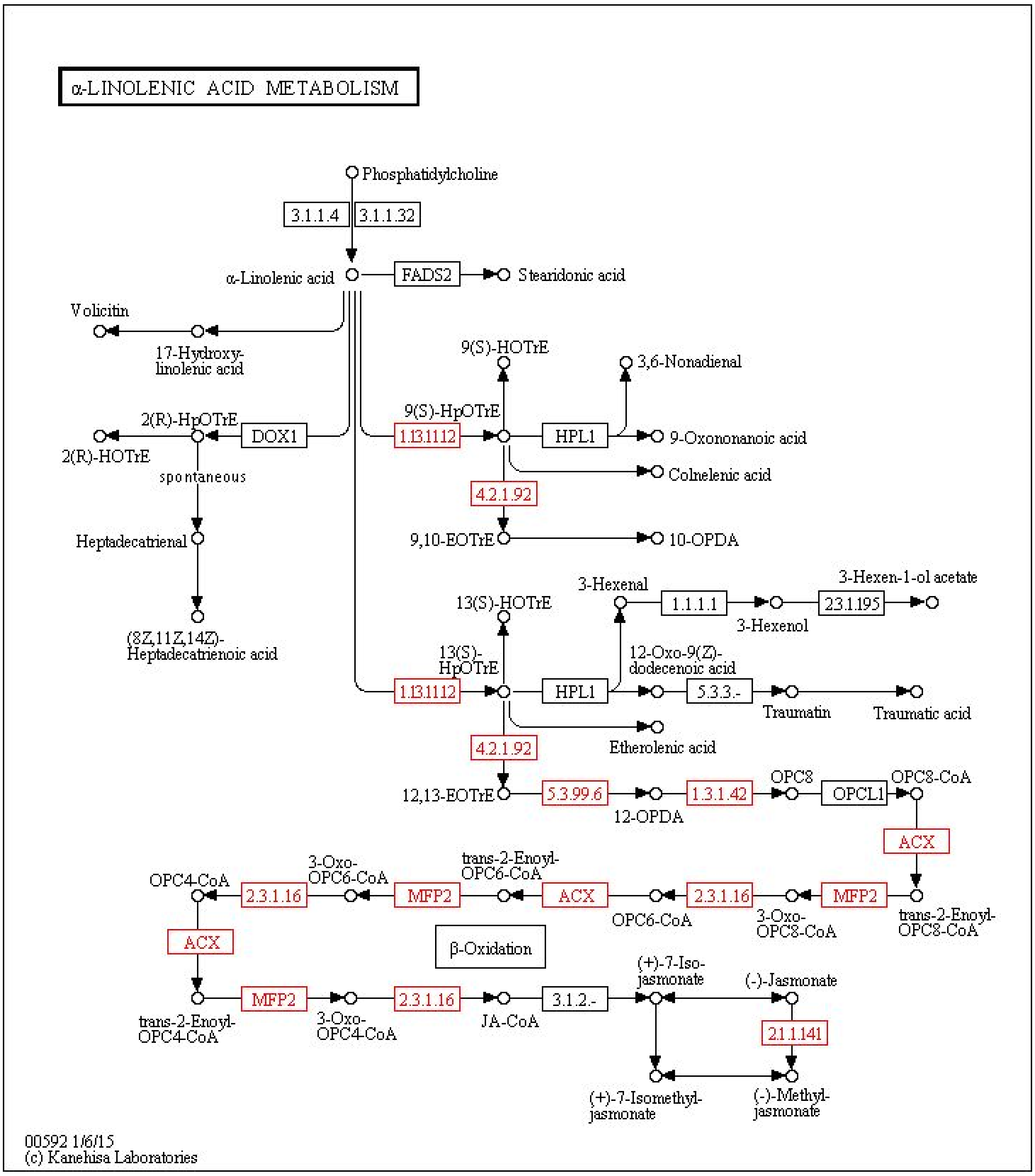
1. Mortazavi,A., Williams,B.A., McCue,K. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5, 621–628.
2. Wang,Z., Gerstein,M., and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.
3. Trapnell,C., Williams,B.A., Pertea,G. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511–515.
4. Garg,R., Patel,R.K., Tyagi,A.K. *et al.* (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.*, 18, 53–63.
5. Wenping,H., Yuan,Z., Jie,S. *et al.* (2011) De novo transcriptome sequencing in Salvia miltiorrhiza to identify genes involved in the biosynthesis of active ingredients. *Genomics*, 98, 272–279.
6. Xia,Z., Xu,H., Zhai,J. *et al.* (2011) RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. *Plant Mol. Biol.*, 77, 299–308.
7. Wang,X.W., Luan,J.B., Li,J.M. *et al.* (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, 11, 400.
8. Schulz,M.H., Zerbino,D.R., Vingron,M. *et al.* (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28, 1086–1092.
9. Grabherr,M.G., Haas,B.J., Yassour,M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29, 644–652.
10. Haas,B.J., Papanicolaou,A., Yassour,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8, 1494–1512.
11. Jones,P., Binns,D., Chang,H.Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236–1240.
12. Lohse,M., Nagel,A., Herter,T. *et al.* (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.*, 37, 1250–1258.
13. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
14. Patro,R., Duggal,G. and Kingsford,C. (2015) Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. *bioRxiv*.
15. Alexa,A. and Rahnenfuhrer,J. (2010) R package version 2, topGO: enrichment analysis for gene ontology. http://www.bioconductor.org/packages/release/bioc/html/topGO.html.
16. Luo,W., Friedman,M.S., Shedden,K. *et al.* (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10, 161.
17. Sanderson,L.A., Ficklin,S.P., Cheng,C.H. *et al.* (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, 2013, bat075.
18. Ficklin,S.P., Sanderson,L.A., Cheng,C.H. *et al.* (2011) Tripal: a construction toolkit for online genome databases. *Database*, 2011, bar044.
19. Smith,R.N., Aleksic,J., Butano,D. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28, 3163–3165.
20. Dietrich,S., Wiegand,S. and Liesegang,H. (2014) TraV: a genome context sensitive transcriptome browser. *PLoS One*, 9, e93677.
21. Nussbaumer,T., Kugler,K.G., Bader,K.C. *et al.* (2014) RNASeqExpressionBrowser—a web interface to browse and visualize high-throughput expression data. *Bioinformatics*, 30, 2519–2520.
22. Vincent,J., Dai,Z., Ravel,C. *et al.* (2013) dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts. *Database*, 2013, bat014.
23. TrinotateWeb: Graphical Interface for Navigating Trinotate Annotations and Expression Analyses. https://trinotate.github.io/TrinotateWeb.html (29 March 2016, date last accessed).
24. Camacho,C., Coulouris,G., Avagyan,V. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
25. Neuhaus,I. CanvasXpress http://canvasxpress.org (29 March 2016, date last accessed).
26. Martin Bland,J. and Altman,D. (1986) Originally published as Volume 1, Issue 8476 Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327, 307–310.
27. Mungall,C.J., Emmert,D.B. and FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23, i337–i346.
28. Fitzpatrick,B. (2004) Distributed caching with Memcached. *Linux J.*, 2004, 5.



Supplementary Material

Supplementary Table 1 - External libraries included in TBro.

Library	Version	Link	Licence	Note
smarty	3.1.13	http://www.smarty.net	LGPL	Server side templating
lightOpenID		http://code.google.com/p/lightopenid	MIT	User authentication
loggedPDO		http://github.com/phryneas/loggedPDO	MIT	Logged Database Connection
Foundation	4.0.8 (js) 4.1.6 (css)	http://foundation.zurb.com/	MIT	Web Framework (css)
jQuery	1.9.1	http://jquery.com	MIT	DOM traversal, event binding, AJAX calls, etc.
jQueryUI	1.10.2	http://jqueryui.com	MIT	autocomplete, accordion, etc.
underscore.js	1.4.4	http://underscorejs.org	MIT	client side templating, helper functions
DateTables	1.9.4	http://www.datatables.net	BSD 3-clause	tables
TableTools	1.0.4	http://datatables.net/extras/tabletools	BSD 3-clause	tables
canvasXpress	7.1	http://canvasxpress.org/	LGPL	plots
sprintf.js		http://github.com/alexei/sprintf.js	BSD 3-clause	
webStorage		https://github.com/ryanttb/webStorage	MIT	local storage (synchronization)
alphanum.js		http://www.davekoelle.com/alphanum.html	LGPL	sorting
PEAR		http://pear.php.net/package	MIT / BSD 2-clause	Log, Console_CommandLine, Console_Table, Console_ProgressBar
Propel		http://propelorm.org	MIT	db abstraction layer
Phing		http://www.phing.info	LGPL	build



Supplementary Figure 1 - KEGG map of the alpha-Linolenic acid metabolism with highlighted components present in a published cart (S1_JA_Pathway). Future releases will color the components dependent on their transcriptional regulation.

Part IV

DISCUSSION

6 GETTING MORE OUT OF BIOLOGICAL DATA

High throughput technologies have revolutionized many fields of biological research. The ever increasing amount of data opens new possibilities and poses challenges at the same time. Algorithms and software tools are required to cope with the data and extract relevant information from it.

Software is vitally important for research but the current system does not support its sustainable development (Bartlett et al., 2017). The best way to solve this problem in the long term is to acknowledge the significance of scientific software and foster its sustainable development. By now, funding agencies are starting to recognize sustainable software as an essential part of research (Deutsche Forschungsgemeinschaft, 2016). Also the Journal of Open Source Software (JOSS) an innovative journal dedicated to software publications has been founded (Smith et al., 2017). This makes it easier to authors of software to get traditional scientific credit in form of publications and citations. Still, large scale change will not happen over night. Thus there are other possibilities to improve quality and sustainability of research software now. There are best practices regarding code style, licensing, testing, and documentation that will facilitate re-use and collaboration by other scientists (Jiménez et al., 2017; List et al., 2017; Prlić and Procter, 2012). Furthermore, training of biologists in basic computation is required (Loman and Watson, 2013; Wilson et al., 2014, 2017). For software designed to be used by others, members of the target community should be involved as early as possible in the development process (Budd et al., 2015). Catalogs like LabWorm (`https://labworm.com/`, accessed 2017/08/31) help scientists find the correct software. A good example for a software designed to make bioinformatic analyses accessible to non-bioinformaticians, reproducible and collaborative is the Galaxy project (Afgan et al., 2016).

The Material and Methods part (page 17) lists numerous measures I took to ensure high quality software. The public availability under an open source license sets the foundation for re-use. Latest technologies have been used in order to gain long term compatibility. Furthermore, all tools are sufficiently developed to be useful as they are. Code quality and documentation allow for both further development and incorporation into other tools.

In the next chapter I discuss how the individual tools developed in this thesis contribute to an advance in their respective fields.

7 ADVANCES IN ECOLOGY, EVOLUTION, AND GENOMICS

7.1 ECOLOGY

In the introduction two challenges in ecology were identified. First the adoption of meta-barcoding as a standard method in pollination ecology (Section 2.1.1). Second the large scale usage of publicly available traits in community ecology (Section 2.1.2).

7.1.1 *Pollen Meta-barcoding*

As outlined in the introduction (Section 2.1.1) adoption of DNA meta-barcoding for pollen is a multi step procedure. One of the prerequisites is selection of a suitable marker. The ITS2 has been validated as a barcode for plants (Chen et al., 2010) and a comprehensive high quality reference database exists (Koetschan et al., 2010, 2012; Schultz et al., 2006; Selig et al., 2008; Wolf et al., 2014). Further, it has been shown that the ITS2 can be used to discriminate species (Müller et al., 2007) despite intragenomic variability (Wolf et al., 2013). Therefore this marker has been selected. To capitalize on recent sequencing efforts I performed an update of the ITS2 database, thereby doubling the amount of sequences (Ankenbrand et al., 2015). After the update 72% of the plant species in the United States of America (USA) are represented in the ITS2 database. An evaluation of pollen meta-barcoding in comparison to traditional methods (Keller et al., 2015) showed its potential. Both the experimental procedure and the bioinformatic processing have been improved to increase efficiency (Sickel et al., 2015). The bioinformatic workflow for data analysis includes a re-training of the classification programs RDP classifier (Wang et al., 2007) and utax (part of usearch, Edgar, 2010). Compared to traditional identification, throughput and assignment depth could be improved. Further, the requirement of expert knowledge for identification has been lowered. The full experimental procedure including bioinformatic processing and classification against a reference database has been compiled into a standard protocol (Sickel et al., in press) and put into context (Keller et al., 2016). This work together with Bell et al. (2016a), Richardson et al. (2015b), and Vere et al. (2015), lay the foundation for DNA meta-barcoding in palynology. The protocol has already been extended by Bell et al. (2017) to use rbcL as an additional marker. Beside plant pollinator interaction research, pollen meta-barcoding has applications in food safety (Bruni et al., 2015), forensics (Bell et al., 2016b), and paleo-climatology (Jørgensen et al., 2012) which greatly benefit from the higher throughput and deeper assignments (Bell et al., 2016a).

7.1.2 *Traits*

The integration of traits into community tables helps address challenges of functional ecology, conservation, and biomonitoring. In microbial community ecology, tools have already been developed to automatically map taxonomy information to functional traits. A common approach is to predict functional capabilities of microbes by mapping the 16S rRNA sequence against a database with fully sequenced and annotated bacteria (Aßhauer et al., 2015; Edgar, 2017; Keller et al., 2014; Langille et al., 2013). For eukaryotic communities this is not yet feasible. Although, all kinds of trait data are available in public databases, no easy way to integrate it automatically for large communities exists. Therefore, I developed the FENNEC a web based workbench that integrates trait data from different sources and makes them readily usable for ecologists (Ankenbrand et al., 2017c). The FENNEC is not intended to be yet another trait database. Instead it holds data from different providers like TraitBank (Parr et al., 2014b) which in turn aggregates data from other providers like TRY (Kattge et al., 2011), LEDA (Kleyer et al., 2008), and IUCN Red List (IUCN, 2017). When a user loads community data into FENNEC, it can be automatically mapped to the organisms in FENNECs database. This way the data can be enriched with every available trait. For broad compatibility with user data I developed a JavaScript library to handle the BIOM format, a standard format for biological observation data (McDonald et al., 2012) together with a corresponding conversion server (Ankenbrand et al., 2017b). This maximizes interoperability with analysis and visualization tools like phyloSeq (McMurdie and Holmes, 2013) and Phinch (Bik and Pitch Interactive, 2014). Unfortunately, both tools accept taxonomy as the only metadata for OTUs. Therefore, FENNEC incorporates a modified version of Phinch able to deal with arbitrary OTU metadata. Additionally, trait data can be exported from FENNEC as pseudo-taxonomy to be accepted by phyloSeq (McMurdie and Holmes, 2013). Other web based tools for community ecology like Qiita (*Qiita* 2016), VAMPS (Huse et al., 2014), or MicrobiomeAnalyst (Dhariwal et al., 2017) provide useful interfaces for analyses but do not use functional trait data. FENNEC lowers the barrier to include traits into ecological community analyses and thus increases its potential to detect biologically relevant signals.

7.2 EVOLUTION

Specialized databases for phylogenetic markers are facing the challenge to keep up with ever increasing data volumes (Section 2.2.1). In addition, the challenge of consistent multi-marker phylogenomics from full bacterial genomes has been described in the introduction (Section 2.2.2).

7.2.1 *ITS2 Phylogeny*

The ITS2 database has been described as an invaluable resource for phylogenetic analyses and barcoding. It adds a layer of quality control and consistency to ITS2 sequences from NCBI (Schultz et al., 2006). In contrast to the raw sequences, the borders are consistently annotated using HMMs (Keller et al., 2009). Further, sequences are validated by their conserved secondary structure (Schultz et al., 2005) and sorted into categories of confidence (Koetschan et al., 2010). However, beside the high quality of the data it is important to keep pace with the new sequences. Data volumes double every view years (Kodama et al., 2012) but phylogenists can only take advantage of this exponential growth if the data is added to the relevant databases. Therefore, the update that led to a duplication of sequences (Ankenbrand et al., 2015) is not only essential for (meta-)barcoding applications (Section 7.1.1) but also for phylogenetic analyses.

7.2.2 *Multi-Marker Trees for Bacteria*

Marker genes like the 16S rRNA gene are commonly used for phylogenetic reconstructions in bacteria (Böttger, 1989; Clarridge, 2004). However the accuracy and robustness of phylogenetic trees can be improved by using multiple markers (Mallo and Posada, 2016). Reasons for inaccuracies in trees from single markers can be stochastic errors, incomplete lineage sorting, gene loss, gene duplication, and horizontal gene transfer (Mallo and Posada, 2016). One of the challenges for multi marker studies is the selection of appropriate markers for the taxonomic group of interest. Other tools use genome composition (CVTree, Zuo and Hao, 2015) or pre-computed alignments of tight genomic clusters (ATGC, Novichkov et al., 2009). In contrast, the Bacterial Core Genome Tree (bcgTree) implements an automated process to build trees from conserved core genes across all bacterial clades (Ankenbrand and Keller, 2016). The set of 107 genes has been shown to be available in at least 95% of full bacterial genomes in a single copy (Dupont et al., 2012). They can additionally be used as a proxy for the completeness of draft assemblies (Albertsen et al., 2013). For a quick overview bcgTree produces an absence presence map of all genes in all genomes. The selection of those genes ensures that a reasonably large fraction contributes to the reconstructed tree while there is little chance that missing genes bias the analysis. Thus facilitating accurate and robust phylogenetic tree reconstruction utilizing larger fractions of the available data. A validation of this method has been performed by comparing resulting trees to the known phylogeny of Lactobacillales and corresponding 16S gene trees. Furtheremore, the robustness of trees has been shown to increase with increasing number of genes. The bcgTree features a command line interface that chains all required steps together conveniently. Additionally, a graphical user interface eases usage by non-bioinformaticians. One of the main applications of bcgTree is the fast, yet accurate, placement of newly sequenced genomes in the tree of life.

7.3 GENOMICS

Genomics is one of the fields in biology with the longest history of open data sharing (Kaye et al., 2009). HTS has accelerated the pace at which this data is generated. The genomic challenges addressed in this thesis include the automatic extraction and assembly of plastid genomes (Section 2.3.1), the interactive visualization of whole genome alignments (Section 2.3.2), and the exploration of *de novo* transcriptomes (Section 2.3.3).

7.3.1 *Plastid Genomes*

Chloroplast genomes of many plants incur as a by product of their genome sequencing projects. Although often not the matter of interest in the original project they can yield valuable insights (Daniell et al., 2016). If chloroplast DNA is not treated specifically it appears with much larger coverage than the rest of the genome. A simple way to get chloroplast sequences is doing a full assembly of the genome e. g. with SPAdes (Bankevich et al., 2012) and use BLAST (Altschul et al., 1990) to compare the resulting contigs against a database of chloroplasts. However, this approach is time consuming and rarely leads to a single contig containing all of the chloroplast genome. A targeted assembly of chloroplast sequences that is aware of the genomic structure, with two single copy regions and an inverted repeat, promises higher chances for success. The chloroExtractor does that by using the k-mer distribution (determined by jellyfish (Marçais and Kingsford, 2011)) and extracting reads from the chloroplast peak. This way complete chloroplasts are assembled from genomic sequencing reads (Section 5.9). Extensions of barcoding to use full chloroplast genomes have been suggested and are used in the PhyloAlps project (Coissac et al., 2016). This group also developed a dedicated organelle assembler, org.asm which is intended to assemble chloroplasts or mitochondria from whole genome sequences but is not yet production ready (Coissac, unpublished). It has been demonstrated that full chloroplast genomes can improve phylogenetic resolution in plants (Huang et al., 2016). The availability of many full chloroplast genomes facilitates comparative genomic analyses. Verdant is a web resource collecting whole chloroplast genome sequences and annotations (McKain et al., 2017). The utility of chloroExtractor comes from the targeted extraction of chloroplasts from whole genome shotgun data. Thus plant genome projects can use it to quickly solve this fraction of the genome. One of the next steps is an automatic mining of Sequence Read Archive (SRA) (Leinonen et al., 2011) to find chloroplast genomes hidden in the pile of data.

7.3.2 *Comparative Genomics*

Comparison of genome architecture between chloroplasts, bacteria or even eukaryotic chromosomes requires whole genome alignment tools. Those tools, e. g. lastz (*LASTZ* 2015), MUMmer 2 (Delcher et al., 2002), or Cactus (Paten et al., 2011) produce textual output consisting of tables indicating which fractions of the sequences correspond to each other. This output is not easy to interpret by direct inspection. Therefore, visualization of the results is essential. However, existing tools like EasyFig (Sullivan et al., 2011) and genoPlotR (Guy et al., 2010) lack the capability to interactively change the visualization. Others (e. g. BRIG (Alikhan et al., 2011)) have a circular layout which does not scale well to compare multiple genomes. The AliTV visualizes multiple WGAs in a linear layout and provides interactive capabilities to dynamically change the produced graphic (Ankenbrand et al., 2017a). In addition to the visualization interface in the web browser AliTV comes with perl scripts that automate the process of calculating the WGAs and preparing the files required for illustration. The interactive features help to correct common artifacts like split links on circular sequences, filter irrelevant information (e. g. low identity links) and zoom into regions of interest. Consequently the AliTV assists in generating hypotheses about differences in genome architecture.

7.3.3 *De-Novo Transcriptomics*

Transcriptomics moves from a genetic inventory to analyzing which genes are actually expressed in given tissues, life stages, and conditions (Wang et al., 2009). There are plenty of tools to produce functional annotations for transcripts e. g. interproScan (Jones et al., 2014), Mercator (Lohse et al., 2014), and blast2go (Conesa and Götz, 2008). Also tools to quantify expression levels are available using mappings (RSEM (Li and Dewey, 2011)) or k-mer counts (sailfish (Patro et al., 2014), salmon (Patro et al., 2015), kallisto (Bray et al., 2016)). Differentially expressed genes can be identified with R (R Core Team, 2017) packages, DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010). However, interpreting results from those tools without command-line knowledge is difficult because annotations, expression counts and differential expression results are usually scattered over multiple text files. Further, there are no easy ways to search e. g. by homology (using BLAST (Altschul et al., 1990)) or to subset the data. If a reference genome is available the gene expression profiles can be added to genome browsers like GBrowse (Stein et al., 2002). This option is not available for *de novo* transcriptomes. TBro bridges the gap between the data and lab biologists by providing an intuitive interface to search, subset, and analyze RNA-Seq datasets (Ankenbrand et al., 2016). TBro was already used to analyze the Venus flytrap transcriptome (Bemm et al., 2016) and still provides an interface to explore it for anyone interested.

Recent research indicates that specific plant genes influence the composition of the leaf microbiome (Brachi et al., 2017). Thus directly linking transcriptomics to community ecology. Beside this biological connection there is also a methodological connection. As data is structured very similarly, both can be stored in BIOM files (McDonald et al., 2012) and tools developed for differential gene expression analyses are used for differential abundance analyses as well (Paulson et al., 2013b).

8 FUTURE WORK

At this point all of the developed tools are ready to be used. However, they are relatively isolated. So it is possible to run the pollen analysis pipeline as described in Sickel et al. (in press) and then import the results into FENNEC (Ankenbrand et al., 2017c). Likewise it is possible to calculate a phylogenetic tree from bacteria found on different flowers using bcgTree and analyze the transcription patterns of those flowers in TBro in parallel. However, the integration of results from the different approaches is not yet well supported. So one of the most beneficial next steps besides extending the basic functionality of each individual tool is improving their interactions. In order to understand complex biological systems an integrative approach combining information from different cellular and organismal levels is mandatory. Furthermore, methods from different fields of biology are required to achieve this goal, including among others physiology, genetics, biochemistry, and phylogeny (Raes and Bork, 2008; Zaneveld et al., 2011). Integration of data from multiple disciplines promises the best chances to unravel large scale ecological phenomena (Fierer et al., 2012; Raes et al., 2011; Roux et al., 2016). An important aspect to investigate is the role of evolution in shaping the interaction between ecological traits and function (Braakman et al., 2017). Facing the challenges of global change requires a fundamental understanding of ecosystem processes for which organismal traits play an essential role (Bozinovic and Pörtner, 2015; Luque et al., 2013; Shade et al., 2012).

For microbial studies a move from marker based meta-barcoding to shotgun meta-genomics has brought many novel insights (Ranjan et al., 2016; Shah et al., 2011). However, for pollen meta-barcoding this is not feasible because of the much larger and more complex genomes of flowering plants (compared to bacteria). It is also not expected to bring comparable benefits as genomes are much more static with less horizontal gene transfer. However, with further falling costs of sequencing and improvements in technology (e. g. longer reads) shotgun meta-genomics of pollen samples might become an option in the future.

Regarding sustainable scientific software development it is essential to raise awareness for the existing problems among scientists and funders. Recognition of the need for specifically trained Research Software Engineers and creation of appropriate incentives are of paramount importance. Also funds to hire professional software developers for training, consulting and coding, or the creation of institutional software engineering units would improve the current situation drastically (Crouch et al., 2013). The collaboration with professional software developers allows scientists to shift from developer to product owner and spend more time thinking about the scientific problems to solve than squashing bugs in the code.

9 CONCLUSION

The methods and software tools developed during this thesis tackle data related challenges in multiple fields of biology. The scope of tools covers facets of ecology, evolution, and genomics. Much effort has been put into making all parts reproducible and reusable. All of the tools are currently in use to answer real biological questions. As Alexander von Humboldt wrote in his letter to Charles Darwin

> “Les ouvrages ne sont bons, qu’autant qu’ils en font naitre de meilleurs”

which translates to “Scientific contributions are of value only if they give rise to better ones” (Humboldt, 1839). In that sense the true value of this work as a contribution to science arises from its current and future usage.

Part V

APPENDIX

A

INDIVIDUAL AUTHOR CONTRIBUTIONS

This chapter describes the detailed individual author contributions for each publication included in the publications chapter (Chapter 5 on page 25). The first table lists the contributions for each part of the respective study, the second table for each figure and table (if present).

Furthermore, a statement is included on page 194 that legal second publication rights for the manuscripts were obtained, where necessary.

ITS2 DATABASE V: TWICE AS MUCH

Author contribution tables for Ankenbrand et al. (2015) in Section 5.1 on page 26.

Table 1: Individual author contributions for each part of Ankenbrand et al. (2015)

PARTICIPATED IN	AUTHOR INITIALS*				
Study Design Methods Development	JS	MW	FF	MJA	
Data Collection	FF	MJA			
Data Analysis & Interpretation	MJA	FF	AK	JS	MW
Manuscript Writing					
Introduction	MJA	FF	AK	JS	MW
Materials & Methods	MJA	FF	AK	JS	MW
Discussion	MJA	FF	AK	JS	MW
First Draft	MJA	FF	AK	JS	MW

* Responsibility decreasing from left to right.

Table 2: Individual author contributions for the figures and tables of Ankenbrand et al. (2015)

FIGURE/TABLE	AUTHOR INITIALS*	
Table 1	MJA	FF
Supplementary File 1	FF	MJA

* Responsibility decreasing from left to right.

Permission for legal second publication has been granted by the publisher with License Numbers 4092481351885 (text) and 4092511085634 (figures and tables).

Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples

Author contribution tables for Keller et al. (2015) in Section 5.2 on page 38.

Table 3: Individual author contributions for each part of Keller et al. (2015)

Participated in	Author initials*	
Study Design Methods Development	AK	ND/SH/ISD
Data Collection	ND	GG/SR
Data Analysis & Interpretation	ND/KvdO/WvdO	MJA/AK
Manuscript Writing		
Introduction	AK	All others
Materials & Methods	ND	All others
Discussion	AK	All others
First Draft	AK	

[*] Responsibility decreasing from left to right.

Table 4: Individual author contributions for the figures and tables of Keller et al. (2015)

Figure/Table	Author initials*	
Figure 1	AK	
Figure 2	AK	
Figure 3	MJA	AK
Figure 4	MJA	AK
Table 1	AK	

[*] Responsibility decreasing from left to right.

Permission for legal second publication has been granted by the publisher with License Number 4092520605723.

INCREASED EFFICIENCY IN IDENTIFYING MIXED POLLEN SAMPLES BY META-BARCODING WITH A DUAL-INDEXING APPROACH

Author contribution tables for Sickel et al. (2015) in Section 5.3 on page 48.

Table 5: Individual author contributions for each part of Sickel et al. (2015)

PARTICIPATED IN	AUTHOR INITIALS*		
Study Design	AK	AH/ISD	JL/SH
Methods Development	WS/GG	AK/MJA	
Data Collection	WS/GG	MJA	JL
Data Analysis & Interpretation	WS/MJA	AK	
Manuscript Writing			
Introduction	WS	AK/MJA	GG/AH/SH/JL/ISD
Materials & Methods	WS	AK/MJA	GG/AH/SH/JL/ISD
Discussion	WS	AK/MJA	GG/AH/SH/JL/ISD
First Draft	WS	AK/MJA	GG/AH/SH/JL/ISD

* Responsibility decreasing from left to right.

Table 6: Individual author contributions for the figures and tables of Sickel et al. (2015)

FIGURE/TABLE	AUTHOR INITIALS*		
Figure 1	WS	AK	
Figure 2	WS	MJA	
Figure 3	WS	AK	
Supplementary File 1	WS	MJA	JL
Supplementary File 2	MJA		
Supplementary File 2	MJA		

* Responsibility decreasing from left to right.

STANDARD METHOD FOR IDENTIFICATION OF BEE POLLEN MIXTURES THROUGH META-BARCODING

Author contribution tables for Sickel et al. (in press) in Section 5.4 on page 70.

Table 7: Individual author contributions for each part of Sickel et al. (in press)

PARTICIPATED IN	AUTHOR INITIALS*			
Study Design	N/A			
Methods Development	WS/MJA	GG/AK	FF	ISD
Data Collection	N/A			
Data Analysis & Interpretation	N/A			
Manuscript Writing				
Introduction	WS/MJA/AK	GG/FF	ISD	
Materials & Methods	WS/MJA/AK	GG/FF	ISD	
Discussion	N/A			
First Draft	WS/MJA/AK	GG/FF	ISD	

* Responsibility decreasing from left to right.

Table 8: Individual author contributions for the figures and tables of Sickel et al. (in press)

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	WS	AK
Figure 2	WS	AK
Table 1	WS	

* Responsibility decreasing from left to right.

DNA-METABARCODING - EIN NEUER BLICK AUF ORGANISMISCHE DIVERSITÄT

Author contribution tables for Keller et al. (2016) in Section 5.5 on page 82.

Table 9: Individual author contributions for each part of Keller et al. (2016)

PARTICIPATED IN	AUTHOR INITIALS*	
Study Design	N/A	
Methods Development	N/A	
Data Collection	N/A	
Data Analysis & Interpretation	N/A	
Manuscript Writing		
Introduction	AK	WS/MJA/GG
Materials & Methods	N/A	
Discussion	AK	WS/MJA/GG
First Draft	AK	

* Responsibility decreasing from left to right.

Table 10: Individual author contributions for the figures and tables of Keller et al. (2016)

FIGURE/TABLE	AUTHOR INITIALS*
Figure 1	AK
Figure 2	WS
Figure 3	AK

* Responsibility decreasing from left to right.

Permission for legal second publication has been granted by the publisher with License Numbers 4092970266068 and 4092970135081.

BIOJS-IO-BIOM, A BIOJS COMPONENT FOR HANDLING DATA IN BIOLOGICAL OBSERVATION MATRIX (BIOM) FROMAT

Author contribution tables for Ankenbrand et al. (2017b) in Section 5.6 on page 86.

Table 11: Individual author contributions for each part of Ankenbrand et al. (2017b)

PARTICIPATED IN	AUTHOR INITIALS*			
Study Design	MJA	NT	FF/AK	SH
Methods Development	MJA	SH		
Data Collection	N/A			
Data Analysis & Interpretation	N/A			
Manuscript Writing				
Introduction	MJA	All others		
Materials & Methods	MJA	All others		
Discussion	MJA	All others		
First Draft	MJA			

* Responsibility decreasing from left to right.

Table 12: Individual author contributions for the figures and tables of Ankenbrand et al. (2017b)

FIGURE/TABLE	AUTHOR INITIALS*
Listing 1	MJA
Listing 2	MJA

* Responsibility decreasing from left to right.

FENNEC – FUNCTIONAL EXPLORATION OF NATURAL NETWORKS AND ECOLOGICAL COMMUNITIES

Author contribution tables for Ankenbrand et al. (2017c) in Section 5.7 on page 101.

Table 13: Individual author contributions for each part of Ankenbrand et al. (2017c)

PARTICIPATED IN	AUTHOR INITIALS*			
Study Design	MJA	AK		
Methods Development	MJA	SH	FF	AK
Data Collection	MJA	AK		
Data Analysis & Interpretation	MJA	AK	SH	
Manuscript Writing				
Introduction	MJA	AK	SH	FF
Materials & Methods	MJA	AK	SH	FF
Discussion MJA	AK	SH	FF	
First Draft	MJA	SH		

* Responsibility decreasing from left to right.

Table 14: Individual author contributions for the figures and tables of Ankenbrand et al. (2017c)

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	MJA	SH
Figure 2	MJA	AK
Figure 3	SH	
Figure 4	MJA	
Table 1	MJA	AK
Supplemental Figure 1	MJA	FF
Supplemental File S1	MJA	
Supplemental File S2	MJA	

* Responsibility decreasing from left to right.

BCGTREE: AUTOMATIZED PHYLOGENETIC TREE BUILDING FROM BACTERIAL CORE GENOMES

Author contribution tables for Ankenbrand and Keller (2016) in Section 5.8 on page 122.

Table 15: Individual author contributions for each part of Ankenbrand and Keller (2016)

PARTICIPATED IN	AUTHOR INITIALS*	
Study Design	AK	
Methods Development	MJA	
Data Collection	MJA	
Data Analysis & Interpretation	MJA/AK	
Manuscript Writing		
Introduction	MJA	AK
Materials & Methods	MJA	AK
Discussion	MJA	AK
First Draft	MJA	AK

* Responsibility decreasing from left to right.

Table 16: Individual author contributions for the figures and tables of Ankenbrand and Keller (2016)

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	MJA	AK
Figure 2	MJA	AK
Figure 3	AK	MJA
Figure 4	AK	MJA
Figure 5	AK	MJA
Supplementary Figure 1	MJA	AK
Supplementary Figure 2	AK	MJA
Supplementary Figure 3	AK	MJA

* Responsibility decreasing from left to right.

Permission for legal second publication has been granted by the publisher with License Number 4092511470118.

CHLOROEXTRACTOR: EXTRACTION AND ASSEMBLY OF THE CHLOROPLAST GENOME FROM WHOLE GENOME SHOTGUN DATA

Author contribution tables for the chloroExtractor manuscript in Section 5.9 on page 135 as submitted to the Journal of Open Source Software (JOSS).

Table 17: Individual author contributions for each part of the chloroExtractor manuscript

PARTICIPATED IN	AUTHOR INITIALS*				
Study Design	TH/FF	MJA			All others
Methods Development	MJA	TH	SP	FF	All others
Data Collection	SP	FF	MQ	TH	All others
Data Analysis & Interpretation	SP	MJA	FF	TH	All others
Manuscript Writing					
Introduction	MJA	FF	SP		All others
Materials & Methods	MJA	FF			All others
Discussion	FF	MJA			All others
First Draft	MJA				

* Responsibility decreasing from left to right.

Table 18: Individual author contributions for the figures and tables of the chloroExtractor manuscript

FIGURE/TABLE	AUTHOR INITIALS*
Figure 1	NT

* Responsibility decreasing from left to right.

ALITV – INTERACTIVE VISUALIZATION OF WHOLE GENOME COMPARISONS

Author contribution tables for Ankenbrand et al. (2017a) in Section 5.10 on page 139.

Table 19: Individual author contributions for each part of Ankenbrand et al. (2017a)

PARTICIPATED IN	AUTHOR INITIALS*			
Study Design	MJA/FF	TH	SH	
Methods Development	MJA/SH	FF	TH	
Data Collection	MJA	FF	SH/TH	
Data Analysis & Interpretation	All authors			
Manuscript Writing				
Introduction	TH	FF	MJA	SH
Materials & Methods	SH	MJA/FF		
Discussion	MJA	FF	TH/SH	
First Draft	MJA			

* Responsibility decreasing from left to right.

Table 20: Individual author contributions for the figures and tables of Ankenbrand et al. (2017a)

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	MJA	
Table 1	SH	MJA

* Responsibility decreasing from left to right.

TBRO: VISUALIZATION AND MANAGEMENT OF *DE NOVO* TRANSCRIPTOMES

Author contribution tables for Ankenbrand et al. (2016) in Section 5.11 on page 156.

Table 21: Individual author contributions for each part of Ankenbrand et al. (2016)

PARTICIPATED IN	AUTHOR INITIALS*			
Study Design	FB	FF	DB	LW/MJA
Methods Development	LW	MJA	FB	
Data Collection	FB			
Data Analysis & Interpretation	FB	MJA	FF/DB	
Manuscript Writing				
Introduction	FB	MJA		All others
Materials & Methods	MJA	FB		All others
Discussion	FB	MJA		All others
First Draft	FB/MJA			

* Responsibility decreasing from left to right.

Table 22: Individual author contributions for the figures and tables of Ankenbrand et al. (2016)

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	FB	MJA
Figure 2	FB	MJA
Supplementary Table 1	MJA	
Supplementary Table 2	FB	

* Responsibility decreasing from left to right.

CONFIRMATION

The doctoral researcher confirms that he has obtained permission from both the publishers and the co-authors for legal second publication, where applicable.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment in this chapter.

Markus Johannes Ankenbrand

Doctoral researcher's name   Date, Place   Signature

PD Dr. Alexander Keller

Primary supervisor's name   Date, Place   Signature

AWARDS

2014–present	Fellow of the Graduate School of Life Sciences *German Excellence Initiative*
2015	Award for excellent bachelor studies and outstanding thesis *Faculty of Informatics, University of Würzburg, Germany*
2012–2014	Deutschlandstipendium *Federal Ministry of Education and Research, Germany*
2012	Award for excellent bachelor thesis *Faculty of Biology, University of Würzburg, Germany*
2008	Siemenspreis for the best natural science university entrance diploma *Siemens Stiftung, Munich, Germany*

POSTERS AND PRESENTATIONS

Presentation

Ankenbrand MJ, Keller A, Koetschan C, Wolf M, Schultz J, Förster F. 2015. *Extending the ITS2-workbench with DNA barcoding capabilities*. 6th International Barcode of Life Conference, Guelph, Canada. DOI:10.5281/zenodo.30543

Selected posters

Ankenbrand MJ, Weber L, Becker D, Förster F, Bemm F. 2016. *TBro - A Transcriptome Browser For De Novo RNA-Sequencing Experiments*. German Conference on Bioinformatics, Berlin, Germany. DOI:10.5281/zenodo.61590

Ankenbrand MJ, Hohlfeld S, Förster F. 2015. *AliTV - Alignment Toolbox and Visualization*. Eureka!, Würzburg, Germany. DOI:10.5281/zenodo.32014

Ankenbrand M, Grimmer G, Härtel S, Steffan-Dewenter I, Keller A. 2014. *Classification of Mixed Plant Samples by Next-Generation Sequencing*. Eureka!, Würzburg, Germany. (awarded 2nd best poster) DOI:10.5281/zenodo.31153

Würzburg, September 2017

Markus Johannes Ankenbrand

BIBLIOGRAPHY

Adl, S. M. et al. (2012). "The revised classification of eukaryotes." *J Eukaryot Microbiol* 59.5, pp. 429–514. DOI: 10.1111/j.1550-7408.2012.00644.x.

Afgan, E. et al. (2016). "The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update." *Nucleic Acids Research* 44.W1, W3–W10. DOI: 10.1093/nar/gkw343.

Agodi, A., M. Barchitta, A. Grillo, and S. Sciacca (2006). "Detection of genetically modified DNA sequences in milk from the Italian market." *International Journal of Hygiene and Environmental Health* 209.1, pp. 81–88. DOI: 10.1016/j.ijheh.2005.08.005.

Ahrenfeldt, J. et al. (2017). "Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods." *BMC Genomics* 18, p. 19. DOI: 10.1186/s12864-016-3407-6.

Albertsen, M. et al. (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." *Nature Biotechnology* 31.6, pp. 533–538. DOI: 10.1038/nbt.2579.

Alexa, A. and J. Rahnenfuhrer (2010). *R Package Version 2, TopGO: Enrichment Analysis for Gene Ontology*.

Alikhan, N.-F., N. K. Petty, N. L. Ben Zakour, and S. A. Beatson (2011). "BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons." eng. *BMC Genomics* 12, p. 402. DOI: 10.1186/1471-2164-12-402.

Altschul, S., W. Gish, W. Miller, E. Meyers, and D. Lipman (1990). "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215.3, pp. 403–410. DOI: 10.1006/jmbi.1990.9999.

Andrews, S (2010). *FastQC A Quality Control tool for High Throughput Sequence Data*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 09/22/2017).

Angly, F. E., C. J. Fields, and G. W. Tyson (2014). "The Bio-Community Perl Toolkit for Microbial Ecology." *Bioinformatics* 30.13, pp. 1926–1927. DOI: 10.1093/bioinformatics/btu130.

Ankenbrand, M. J. (2016). *Biom-Conversion-Server: Version 1.0.2*. doi: 10.5281/zenodo.218396. DOI: 10.5281/zenodo.218396.

Ankenbrand, M. J. and A. Keller (2016). "bcgTree: Automatized Phylogenetic Tree Building from Bacterial Core Genomes." en. *Genome*. DOI: 10.1139/gen-2015-0175.

Ankenbrand, M. J., A. Keller, M. Wolf, J. Schultz, and F. Förster (2015). "ITS2 Database V: Twice as Much." en. *Molecular Biology and Evolution* 32.11, pp. 3030–3032. DOI: 10.1093/molbev/msv174.

Ankenbrand, M. J., L. Weber, D. Becker, F. Förster, and F. Bemm (2016). "TBro: Visualization and Management of de Novo Transcriptomes." en. *Database* 2016, baw146. DOI: 10.1093/database/baw146.

Ankenbrand, M. J., S. Hohlfeld, T. Hackl, and F. Förster (2017a). "AliTV—interactive Visualization of Whole Genome Comparisons." en. *PeerJ Computer Science* 3, e116. DOI: 10.7717/peerj-cs.116.

Ankenbrand, M. J., N. Terhoeven, S. Hohlfeld, F. Förster, and A. Keller (2017b). "Biojs-Io-Biom, a BioJS Component for Handling Data in Biological Observation Matrix (BIOM) Format." en. *F1000Research* 5, p. 2348. DOI: 10.12688/f1000research.9618.2.

Ankenbrand, M. J., S. C. Y. Hohlfeld, F. Förster, and A. Keller (2017c). "FENNEC – Functional Exploration of Natural Networks and Ecological Communities." en. *bioRxiv*. DOI: 10.1101/194308. eprint: https://www.biorxiv.org/content/early/2017/09/27/194308.full.pdf.

Antipov, D. et al. (2016). "plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data." *bioRxiv*. DOI: 10.1101/048942. eprint: https://www.biorxiv.org/content/early/2016/04/20/048942.full.pdf.

Aronesty, E. (2011). *ea-utils: "Command-line tools for processing biological sequencing data"*. URL: https://github.com/ExpressionAnalysis/ea-utils.

Aronesty, E. (2013). "Comparison of Sequencing Utility Programs." *The Open Bioinformatics Journal* 7.1.

Avrani, S., O. Wurtzel, I. Sharon, R. Sorek, and D. Lindell (2011). "Genomic island variability facilitates Prochlorococcus–virus coexistence." *Nature* 474.7353, pp. 604–608. DOI: 10.1038/nature10172.

Aßhauer, K. P., B. Wemheuer, R. Daniel, and P. Meinicke (2015). "Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data." *Bioinformatics* 31.17, pp. 2882–2884. DOI: 10.1093/bioinformatics/btv287.

Bagella, S. et al. (2013). "Effects of plant community composition and flowering phenology on honeybee foraging in Mediterranean sylvo-pastoral systems." *Applied Vegetation Science* 16.4, pp. 689–697. DOI: 10.1111/avsc.12023.

Baldwin, B. G. (1992). "Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae." *Mol Phylogenet Evol* 1.1, pp. 3–16. DOI: 10.1016/1055-7903(92)90030-K.

Bankevich, A. et al. (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021.

Bartlett, A., B. Penders, and J. Lewis (2017). "Bioinformatics: indispensable, yet hidden in plain sight?" *BMC Bioinformatics* 18, p. 311. DOI: 10.1186/s12859-017-1730-9.

Behling, H., V. Pillar, L. Orlóci, and S. Bauermann (2004). "Late Quaternary Araucaria forest, grassland (Campos), fire and climate dynamics, studied by high-resolution pollen, charcoal and multivariate analysis of the Cambará do Sul core in southern Brazil." *Palaeogeography, Palaeoclimatology, Palaeoecology* 203.3-4, pp. 277–297. DOI: 10.1016/S0031-0182(03)00687-4.

Behm, F., K. Von Der Ohe, and W. Henrich (1996). "Zuverlässigkeit der pollenanalyse von honig bestimmung der pollenhäufigkeit." *Deutsche Lebensmittel-Rundschau* 92.6, pp. 183–188.

Beil, M., H. Horn, and A. Schwabe (2008). "Analysis of pollen loads in a wild bee community (Hymenoptera: Apidae) - A method for elucidating habitat use and foraging distances." *Apidologie* 39.4, pp. 456–467. DOI: 10.1051/apido:2008021.

Bell, K. L. et al. (2016a). "Pollen DNA barcoding: current applications and future prospects." *Genome* 59.9, pp. 629–640. DOI: 10.1139/gen-2015-0200.

Bell, K. L., K. S. Burgess, K. C. Okamoto, R. Aranda, and B. J. Brosi (2016b). "Review and future prospects for DNA barcoding methods in forensic palynology." *Forensic Science International: Genetics* 21, pp. 110–116. DOI: 10.1016/j.fsigen.2015.12.010.

Bell, K. L. et al. (2017). "Applying pollen DNA metabarcoding to the study of plant–pollinator interactions1." *Applications in Plant Sciences* 5.6. DOI: 10.3732/apps.1600124.

Bemm, F. et al. (2016). "Venus Flytrap Carnivorous Lifestyle Builds on Herbivore Defense Strategies." en. *Genome Research*. DOI: 10.1101/gr.202200.115.

Bennett, K. and L. Parducci (2006). "DNA from pollen: Principles and potential." *Holocene* 16.8, pp. 1031–1034. DOI: 10.1177/0959683606069383.

Benson, D., I. Karsch-Mizrachi, D. Lipman, J. Ostell, and E. Sayers (2009). "GenBank." *Nucleic Acids Research* 38.SUPPL.1, pp. D46–D51. DOI: 10.1093/nar/gkp1024.

Benson, D. A. et al. (2013). "GenBank." eng. *Nucleic Acids Res* 41.Database issue, pp. D36–D42. DOI: 10.1093/nar/gks1195.

Bijlmakers, H. (2017). *World Crops Database - Fruits, vegetables, cereals and other agricultural crops*. World Crops Database. URL: http://world-crops.com/ (visited on 07/25/2017).

Bik, H. M. and Pitch Interactive (2014). "Phinch: An Interactive, Exploratory Data Visualization Framework for –Omic Datasets." en. *bioRxiv*, p. 009944. DOI: 10.1101/009944.

Biom Javascript Module · Issue #699 · Biocore/Biom-Format. https://github.com/biocore/biom-format/issues/699. Accessed: 2016-09-08.

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics (Oxford, England)* 30.15, pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bork, P. (2005). "Is there biological research beyond Systems Biology? A comparative analysis of terms." *Molecular Systems Biology* 1, p. 2005.0012. DOI: 10.1038/msb4100016.

Bosch, J., A. Martín González, A. Rodrigo, and D. Navarro (2009). "Plant-pollinator networks: Adding the pollinator's perspective." *Ecology Letters* 12.5, pp. 409–419. DOI: 10.1111/j.1461-0248.2009.01296.x.

Botanischer Informationsknoten Bayern, Germany. "Staatliche Naturwissenschaftliche Sammlungen Bayern."

Bozinovic, F. and H.-O. Pörtner (2015). "Physiological ecology meets climate change." *Ecology and Evolution* 5.5, pp. 1025–1030. DOI: 10.1002/ece3.1403.

Braakman, R., M. J. Follows, and S. W. Chisholm (2017). "Metabolic evolution and the self-organization of ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 114.15, E3091–E3100. DOI: 10.1073/pnas.1619573114.

Brachi, B. et al. (2017). "Plant genes influence microbial hubs that shape beneficial leaf communities." *bioRxiv*, p. 181198. DOI: 10.1101/181198.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter (2016). "Near-optimal probabilistic RNA-seq quantification." *Nature Biotechnology* 34.5, pp. 525–527. DOI: 10.1038/nbt.3519.

Brazma, A. et al. (2001). "Minimum Information about a Microarray Experiment (MIAME)—toward Standards for Microarray Data." en. *Nature Genetics* 29.4, pp. 365–371. DOI: 10.1038/ng1201-365.

Bridge, P., P. Roberts, B. Spooner, and G. Panchal (2003). "On the unreliability of published DNA sequences." *New Phytologist* 160.1, pp. 43–48. DOI: 10.1046/j.1469-8137.2003.00861.x.

Brown, C. T. (2015a). *Is software a primary product of science?* URL: http://ivory.idyll.org/blog/2015-software-as-a-primary-product-of-science.html (visited on 07/11/2017).

Brown, C. T. (2015b). *More on scientific software*. URL: http://ivory.idyll.org/blog/2015-more-on-software.html (visited on 07/11/2017).

Bruni, I. et al. (2015). "A DNA barcoding approach to identify plant species in multiflower honey." *Food Chemistry* 170, pp. 308–315. DOI: 10.1016/j.foodchem.2014.08.060.

Buchheim, M. et al. (2011). "Internal transcribed spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: Towards an automated reconstruction of the green algal tree of life." *PLoS ONE* 6.2. DOI: 10.1371/journal.pone.0016931.

Budd, A. et al. (2015). "A Quick Guide for Building a Successful Bioinformatics Community." *PLOS Computational Biology* 11.2, e1003972. DOI: 10.1371/journal.pcbi.1003972.

Budrys, E, A Budriene, and S Orlovskyte (2014). *Cavity-nesting wasps and bees database*. URL: http://scales.ckff.si/scaletool/?menu=6&submenu=3 (visited on 09/22/2017).

Burrows, P. A. (1998). "Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes." *The EMBO Journal* 17.4, pp. 868–876. DOI: 10.1093/emboj/17.4.868.

Butler, D. and P. Smaglik (2000). "Draft data leave geneticists with a mountain still to climb." *Nature* 405.6790, pp. 984–985. DOI: 10.1038/35016703.

Bálint, M., P.-A. Schmidt, R. Sharma, M. Thines, and I. Schmitt (2014). "An Illumina metabarcoding pipeline for fungi." *Ecology and Evolution* 4.13, pp. 2642–2653. DOI: 10.1002/ece3.1107.

Böttger, E. C. (1989). "Rapid determination of bacterial ribosomal RNA sequences by direct sequencing of enzymatically amplified DNA." *FEMS microbiology letters* 53.1, pp. 171–176.

Camacho, C. et al. (2009). "BLAST+: Architecture and applications." *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-421.

Cantarel, B. L. et al. (2008). "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes." *Genome Research* 18.1, pp. 188–196. DOI: 10.1101/gr.6743907.

Capella-Gutierrez, S., F. Kauff, and T. Gabaldón (2014). "A phylogenomics approach for selecting robust sets of phylogenetic markers." *Nucleic Acids Research* 42.7, e54. DOI: 10.1093/nar/gku071.

Caporaso, J. G. et al. (2010). "QIIME Allows Analysis of High-Throughput Community Sequencing Data." en. *Nature Methods* 7.5, pp. 335–336. DOI: 10.1038/nmeth.f.303.

Carvalho, B. S. and G. Rustici (2013). "The challenges of delivering bioinformatics training in the analysis of high-throughput data." *Briefings in Bioinformatics* 14.5, pp. 538–547. DOI: 10.1093/bib/bbt018.

Carvell, C., P. Westrich, W. Meek, R. Pywell, and M. Nowakowski (2006). "Assessing the value of annual and perennial forage mixtures for bumblebees by direct observation and pollen analysis." *Apidologie* 37.3, pp. 326–340. DOI: 10.1051/apido:2006002.

Carver, T. et al. (2008). "Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database." *Bioinformatics (Oxford, England)* 24.23, pp. 2672–6. DOI: 10.1093/bioinformatics/btn529.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." *Molecular Biology and Evolution* 17.4, pp. 540–552.

Cavalier-Smith, T. (1993). "Kingdom protozoa and its 18 phyla." *Microbiological Reviews* 57.4, pp. 953–994.

Chamberlain, S., Z. Foster, I. Bartomeus, D. LeBauer, and D. Harris (2016). *traits: Species Trait Data from Around the Web.*

Chen, S. et al. (2010). "Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species." *PLoS ONE* 5.1. DOI: 10.1371/journal.pone.0008613.

Chin, C.-S. et al. (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." *Nature Methods* may, pp. 1–9. DOI: 10.1038/nmeth.2474.

Ciccarelli, F. et al. (2006). "Toward automatic reconstruction of a highly resolved tree of life." *Science* 311.5765, pp. 1283–1287. DOI: 10.1126/science.1123061.

Clarridge, J. E. (2004). "Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases." *Clinical Microbiology Reviews* 17.4, pp. 840–862. DOI: 10.1128/CMR.17.4.840-862.2004.

Coissac, E., T. Riaz, and N. Puillandre (2012). "Bioinformatic challenges for DNA metabarcoding of plants and animals." *Molecular Ecology* 21.8, pp. 1834–1847. DOI: 10.1111/j.1365-294X.2012.05550.x.

Coissac, E., P. M. Hollingsworth, S. Lavergne, and P. Taberlet (2016). "From barcodes to genomes: extending the concept of DNA barcoding." *Molecular Ecology* 25.7, pp. 1423–1428. DOI: 10.1111/mec.13549.

Cole, J. R. et al. (2005). "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis." *Nucleic Acids Research* 33 (Database Issue), pp. D294–D296. DOI: 10.1093/nar/gki038.

Coleman, A. W. (2003). "ITS2 is a double-edged tool for eukaryote evolutionary comparisons." *Trends Genet* 19.7, pp. 370–375. DOI: 10.1016/S0168-9525(03)00118-5.

Coleman, A. W. (2009). "Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide." *Mol Phylogenet Evol* 50.1, pp. 197–203. DOI: http://dx.doi.org/10.1016/j.ympev.2008.10.008.

Coleman, M. L. et al. (2006). "Genomic islands and the ecology and evolution of Prochlorococcus." *Science (New York, N.Y.)* 311.5768, pp. 1768–70. DOI: 10.1126/science.1122050.

Conesa, A. and S. Götz (2008). "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics." *International Journal of Plant Genomics* 2008. DOI: 10.1155/2008/619832.

Corpas, M. (2014). "The BioJS Article Collection of Open Source Components for Biological Data Visualisation." en. *F1000Research*. DOI: 10.12688/f1000research.3-56.v1.

Corpas, M. et al. (2014). "BioJS: An Open Source Standard for Biological Visualisation –Its Status in 2014." en. *F1000Research*. DOI: 10.12688/f1000research.3-55.v1.

Costa, F. F. (2014). "Big data in biomedicine." *Drug Discovery Today* 19.4, pp. 433–440. DOI: 10.1016/j.drudis.2013.10.012.

Couronne, O. et al. (2003). "Strategies and Tools for Whole-Genome Alignments." *Genome Research* 13.1, pp. 73–80. DOI: 10.1101/gr.762503.

Cowart, D. A. et al. (2015). "Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities." *PLoS ONE* 10.2. DOI: 10.1371/journal.pone.0117562.

Crouch, S. et al. (2013). "The Software Sustainability Institute: Changing Research Software Attitudes and Practices." *Computing in Science Engineering* 15.6, pp. 74–80. DOI: 10.1109/MCSE.2013.133.

Cui, L. et al. (2006). "ChloroplastDB: the Chloroplast Genome Database." *Nucleic Acids Research* 34 (Database issue), pp. D692–696. DOI: 10.1093/nar/gkj055.

Daniell, H., C.-S. Lin, M. Yu, and W.-J. Chang (2016). "Chloroplast genomes: diversity, evolution, and applications in genetic engineering." *Genome Biology* 17. DOI: 10.1186/s13059-016-1004-2.

Darling, A. C. E., B. Mau, F. R. Blattner, and N. T. Perna (2004). "Mauve: multiple alignment of conserved genomic sequence with rearrangements." eng. *Genome Res* 14.7, pp. 1394–1403. DOI: 10.1101/gr.2289704.

Darling, A. E., B. Mau, and N. T. Perna (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." eng. *PLoS One* 5.6, e11147. DOI: 10.1371/journal.pone.0011147.

Darling, J. A. and M. J. Blum (2007). "DNA-based methods for monitoring invasive species: a review and prospectus." *Biological Invasions* 9.7, pp. 751–765. DOI: 10.1007/s10530-006-9079-4.

Davies, A. and R. Tipping (2004). "Sensing small-scale human activity in the palaeoecological record: Fine spatial resolution pollen analyses from Glen Affric, northern Scotland." *Holocene* 14.2, pp. 233–245. DOI: 10.1191/0959683604hl701rp.

Davies, C. and P. Fall (2001). "Modern pollen precipitation from an elevational transect in central Jordan and its relationship to vegetation." *Journal of Biogeography* 28.10, pp. 1195–1210. DOI: 10.1046/j.1365-2699.2001.00630.x.

Delcher, A. L., A. Phillippy, J. Carlton, and S. L. Salzberg (2002). "Fast algorithms for large-scale genome alignment and comparison." *Nucleic Acids Research* 30.11, pp. 2478–2483.

Deutsche Forschungsgemeinschaft (2016). *Ausschreibung: Nachhaltigkeit von Forschungssoftware*.

Dhariwal, A. et al. (2017). "MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data." *Nucleic Acids Research* 45 (W1), W180–W188. DOI: 10.1093/nar/gkx295.

Didelot, X, G Méric, and D Falush (2012). "Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli." *BMC*.

Dietrich, S., S. Wiegand, and H. Liesegang (2014). "TraV: A genome context sensitive transcriptome browser." *PLoS ONE* 9.4. DOI: 10.1371/journal.pone.0093677.

Dijk, E. L. v., H. Auger, Y. Jaszczyszyn, and C. Thermes (2014). "Ten years of next-generation sequencing technology." *Trends in Genetics* 30.9, pp. 418–426. DOI: 10.1016/j.tig.2014.07.001.

Dixon, P. (2003). "VEGAN, a package of R functions for community ecology." *Journal of Vegetation Science* 14.6, pp. 927–930.

Dolinski, K. and O. G. Troyanskaya (2015). "Implications of Big Data for cell biology." *Molecular Biology of the Cell* 26.14, pp. 2575–2578. DOI: 10.1091/mbc.E13-12-0756.

Dupont, C. et al. (2012). "Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage." *ISME Journal* 6.6, pp. 1186–1199. DOI: 10.1038/ismej.2011.189.

Dupuis, J. R., A. D. Roe, and F. A. H. Sperling (2012). "Multi-locus species delimitation in closely related animals and fungi: one marker is not enough." *Molecular Ecology* 21.18, pp. 4422–4436. DOI: 10.1111/j.1365-294X.2012.05642.x.

EPPO (2017). *EPPO Global Database (available online)*. URL: https://gd.eppo.int/ (visited on 09/22/2017).

Eddy, S. (2010). *HMMER3: A New Generation of Sequence Homology Search Software*.

Edgar, R. (2004). "MUSCLE: Multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research* 32.5, pp. 1792–1797. DOI: 10.1093/nar/gkh340.

Edgar, R. (2010). "Search and clustering orders of magnitude faster than BLAST." *Bioinformatics* 26.19, pp. 2460–2461. DOI: 10.1093/bioinformatics/btq461.

Edgar, R. (2013). "UPARSE: Highly accurate OTU sequences from microbial amplicon reads." *Nature Methods* 10.10, pp. 996–998. DOI: 10.1038/nmeth.2604.

Edgar, R. C. (2017). "SINAPS: Prediction of microbial traits from marker gene sequences." *bioRxiv*, p. 124156. DOI: 10.1101/124156.

Eklund, A., T. E. Nichols, and H. Knutsson (2016). "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates." *Proceedings of the National Academy of Sciences* 113.28, pp. 7900–7905. DOI: 10.1073/pnas.1602413113.

Elbrecht, V. and F. Leese (2015). "Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol." *PLOS ONE* 10.7, e0130324. DOI: 10.1371/journal.pone.0130324.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller (2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* 32.19, pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

Federhen, S. (2012). "The NCBI Taxonomy database." *Nucleic Acids Res* 40.Database issue, pp. D136–D143. DOI: 10.1093/nar/gkr1178.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." *Journal of Molecular Evolution* 17.6, pp. 368–376.

Felsenstein, J. (1989). "PHYLIP - Phylogeny Inference Package (Version 3.2)." *Cladistics* 5, pp. 164–166.

Ficklin, S. et al. (2011). "Tripal: A construction toolkit for online genome databases." *Database* 2011. DOI: 10.1093/database/bar044.

Field, D. et al. (2011). "The Genomic Standards Consortium." *PLOS Biology* 9.6, e1001088. DOI: 10.1371/journal.pbio.1001088.

Fierer, N., M. Hamady, C. Lauber, and R. Knight (2008). "The influence of sex, handedness, and washing on the diversity of hand surface bacteria." *Proceedings of the National Academy of Sciences of the United States of America* 105.46, pp. 17994–17999. DOI: 10.1073/pnas.0807920105.

Fierer, N. et al. (2012). "Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients." *The ISME journal* 6.5, pp. 1007–1017. DOI: 10.1038/ismej.2011.159.

Finn, R. et al. (2009). "The Pfam protein families database." *Nucleic Acids Research* 38.SUPPL.1, pp. D211–D222. DOI: 10.1093/nar/gkp985.

Fischer, M. G. (2015). "Virophages go nuclear in the marine alga <i>Bigelowiella natans</i>." *Proceedings of the National Academy of Sciences* 112.38, pp. 11750–11751. DOI: 10.1073/pnas.1515142112.

Fitch, W. M. (1971). "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology." *Systematic Biology* 20.4, pp. 406–416. DOI: 10.1093/sysbio/20.4.406.

Fitzpatrick, B. (2004). "Distributed caching with memcached." *Linux Journal* 2004.124.

Forsman, A. (2015). "Rethinking phenotypic plasticity and its consequences for individuals, populations and species." *Heredity* 115.4, pp. 276–284. DOI: 10.1038/hdy.2014.92.

Foster, Z. S. L., T. J. Sharpton, and N. J. Grünwald (2017). "Metacoder: An R package for visualization and manipulation of community taxonomic diversity data." *PLOS Computational Biology* 13.2, e1005404. DOI: 10.1371/journal.pcbi.1005404.

Fox, G. E., K. R. Pechman, and C. R. Woese (1977). "Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics." *International Journal of Systematic and Evolutionary Microbiology* 27.1, pp. 44–57. DOI: 10.1099/00207713-27-1-44.

Förster, F. (2015). *NCBI-Taxonomy: First GitHub release v0.80.0*. DOI: 10.5281/zenodo.17383.

Galili, T. (2015). "dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering." *Bioinformatics* 31.22, pp. 3718–3720. DOI: 10.1093/bioinformatics/btv428.

Galimberti, A. et al. (2014). "A DNA Barcoding Approach to Characterize Pollen Collected by Honeybees." *PLOS ONE* 9.10, e109363. DOI: 10.1371/journal.pone.0109363.

Garg, R., R. Patel, A. Tyagi, and M. Jain (2011). "De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification." *DNA Research* 18.1, pp. 53–63. DOI: 10.1093/dnares/dsq028.

Garibaldi, L. A. et al. (2013). "Wild Pollinators Enhance Fruit Set of Crops Regardless of Honey Bee Abundance." *Science* 339.6127, pp. 1608–1611. DOI: 10.1126/science.1230200. eprint: http://science.sciencemag.org/content/339/6127/1608.full.pdf.

Gathmann, A. and T. Tscharntke (2002). "Foraging ranges of solitary bees." *Journal of Animal Ecology* 71.5, pp. 757–764. DOI: 10.1046/j.1365-2656.2002.00641.x.

Gilliam, M. (1990). *Chalkbrood disease of honey bees, Apis mellifera, caused by the fungus, Ascosphaera apis: a review of past and current research*, pp. 398–402.

Girard, M., M. Chagnon, and V. Fournier (2012). "Pollen diversity collected by honey bees in the vicinity of Vaccinium spp. crops and its importance for colony development." *Botany* 90.7, pp. 545–555. DOI: 10.1139/B2012-049.

Goble, C. (2014). "Better Software, Better Research." *IEEE Internet Computing* 18.5, pp. 4–8. DOI: 10.1109/MIC.2014.88.

Grabherr, M. et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nature Biotechnology* 29.7, pp. 644–652. DOI: 10.1038/nbt.1883.

Grant, J. R. et al. (2012). "Comparing thousands of circular genomes using the CGView Comparison Tool." *BMC Genomics* 13.1, p. 202. DOI: 10.1186/1471-2164-13-202.

Greene, C. S., J. Tan, M. Ung, J. H. Moore, and C. Cheng (2014). "Big Data Bioinformatics." *Journal of Cellular Physiology* 229.12, pp. 1896–1900. DOI: 10.1002/jcp.24662.

Gugerli, F., L. Parducci, and R. Petit (2005). "Ancient plant DNA: Review and prospects." *New Phytologist* 166.2, pp. 409–418. DOI: 10.1111/j.1469-8137.2005.01360.x.

Guy, L., J. R. Kultima, and S. G. E. Andersson (2010). "genoPlotR: comparative gene and genome visualization in R." eng. *Bioinformatics* 26.18, pp. 2334–2335. DOI: 10.1093/bioinformatics/btq413.

Haas, B. et al. (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." *Nature Protocols* 8.8, pp. 1494–1512. DOI: 10.1038/nprot.2013.084.

Hackl, T., R. Hedrich, J. Schultz, and F. Förster (2014). "proovread: large-scale high-accuracy PacBio correction through iterative short read consensus." *Bioinformatics* 30.21, pp. 3004–3011. DOI: 10.1093/bioinformatics/btu392.

Haft, D., J. Selengut, and O. White (2003). "The TIGRFAMs database of protein families." *Nucleic Acids Research* 31.1, pp. 371–373. DOI: 10.1093/nar/gkg128.

Hajibabaei, M., G. A. C. Singer, P. D. N. Hebert, and D. A. Hickey (2007). "DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics." *Trends in Genetics* 23.4, pp. 167–172. DOI: 10.1016/j.tig.2007.02.001.

Hallin, P. F., T. T. Binnewies, and D. W. Ussery (2008). "The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology." *Molecular bioSystems* 4.5, pp. 363–371. DOI: 10.1039/b717118h.

Han, J.-P., L.-C. Shi, X.-C. Chen, and Y.-L. Lin (2012). "Comparison of four DNA barcodes in identifying certain medicinal plants of Lamiaceae." *Journal of Systematics and Evolution* 50.3, pp. 227–234. DOI: 10.1111/j.1759-6831.2012.00184.x.

Hanley, N., T. D. Breeze, C. Ellis, and D. Goulson (2015). "Measuring the economic value of pollination services: Principles, evidence and knowledge gaps." *Ecosystem Services* 14, pp. 124–132. DOI: 10.1016/j.ecoser.2014.09.013.

Harpke, D. and A. Peterson (2008). "5.8S motifs for the identification of pseudogenic ITS regions." *Botany* 86.3, pp. 300–305. DOI: 10.1139/B07-134.

Harris, R. S. (2007). "Improved pairwise alignment of genomic DNA." PhD thesis. Pennsylvania State University.

Hartfield, M., C. L. Murall, and S. Alizon (2014). "Clinical applications of pathogen phylogenies." *Trends in Molecular Medicine* 20.7, pp. 394–404. DOI: 10.1016/j.molmed.2014.04.002.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard (2003). "Biological identifications through DNA barcodes." *Proceedings. Biological Sciences* 270.1512, pp. 313–321. DOI: 10.1098/rspb.2002.2218.

Hemmer, W. (1997). "Foods derived from genetically modified organisms and detection methods." *Foods Derived from Genetically Modified Organisms and Detection Methods*.

Henle, K. et al. (2014). "Scaling in Ecology and Biodiversity Conservation." *Advanced Books* 1, e1169. DOI: 10.3897/ab.e1169.

Herbig, A, G Jäger, F Battke, and K Nieselt (2012). "GenomeRing: alignment visualization based on SuperGenome coordinates." *Bioinformatics (Oxford, England)* 28.12, pp. i7–15. DOI: 10.1093/bioinformatics/bts217.

Hey, T., S. Tansley, and K. Tolle (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*.

Higgins, D. G. and P. M. Sharp (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." *Gene* 73.1, pp. 237–244.

Hillis, D. M. (1987). "Molecular Versus Morphological Approaches to Systematics." *Annual Review of Ecology and Systematics* 18.1, pp. 23–42. DOI: 10.1146/annurev.es.18.110187.000323.

Hollingsworth, P., S. Graham, and D. Little (2011). "Choosing and using a plant DNA barcode." *PLoS ONE* 6.5. DOI: 10.1371/journal.pone.0019254.

Holt, B. G. and K. A. Jønsson (2014). "Reconciling Hierarchical Taxonomy with Molecular Phylogenies." *Systematic Biology* 63.6, pp. 1010–1017. DOI: 10.1093/sysbio/syu061.

Huang, Y. et al. (2016). "Analysis of Complete Chloroplast Genome Sequences Improves Phylogenetic Resolution in Paris (Melanthiaceae)." *Frontiers in Plant Science* 7. DOI: 10.3389/fpls.2016.01797.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." *Bioinformatics (Oxford, England)* 17.8, pp. 754–755.

Humboldt, A. von (1839). *Letter no. 534*. Darwin Correspondence Project. URL: https://www.darwinproject.ac.uk/letter (visited on 08/31/2017).

Huse, S. M. et al. (2014). "VAMPS: A Website for Visualization and Analysis of Microbial Population Structures." *BMC Bioinformatics* 15, p. 41. DOI: 10.1186/1471-2105-15-41.

Huson, D. H. et al. (2016). "MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data." *PLOS Computational Biology* 12.6, e1004957. DOI: 10.1371/journal.pcbi.1004957.

IUCN (2017). *IUCN Red List of Threatened Species. Version 2017-1*. URL: http://www.iucnredlist.org/ (visited on 09/22/2017).

Ibsen-Jensen, R., K. Chatterjee, and M. A. Nowak (2015). "Computational complexity of ecological and evolutionary spatial dynamics." *Proceedings of the National Academy of Sciences of the United States of America* 112.51, pp. 15636–15641. DOI: 10.1073/pnas.1511366112.

Iliopoulos, I. et al. (2001). "Genome sequences and great expectations." *Genome Biology* 2.1, interactions0001.1–interactions0001.3.

Irschick, D. J. et al. (2013). "Functional ecology: integrative research in the modern age of ecology." *Functional Ecology* 27.1, pp. 1–4. DOI: 10.1111/1365-2435.12037.

Jaspers, E. and J. Overmann (2004). "Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies." *Applied and Environmental Microbiology* 70.8, pp. 4831–4839. DOI: 10.1128/AEM.70.8.4831-4839.2004.

Jiménez, R. C. et al. (2017). "Four simple recommendations to encourage best practices in research software." *F1000Research* 6, p. 876. DOI: 10.12688/f1000research.11407.1.

Jones, P. et al. (2014). "InterProScan 5: Genome-scale protein function classification." *Bioinformatics* 30.9, pp. 1236–1240. DOI: 10.1093/bioinformatics/btu031.

Junker, R. R. and A. Keller (2015). "Microhabitat heterogeneity across leaves and flower organs promotes bacterial diversity." *FEMS microbiology ecology* 91.9, fiv097. DOI: 10.1093/femsec/fiv097.

Junker, R. R., N. Blüthgen, and A. Keller (2015). "Functional and phylogenetic diversity of plant communities differently affect the structure of flower-visitor interactions and reveal convergences in floral traits." *Evolutionary Ecology* 29.3, pp. 437–450. DOI: 10.1007/s10682-014-9747-2.

Junker, R. R., J. Kuppler, A. C. Bathke, M. L. Schreyer, and W. Trutschnig (2016). "Dynamic range boxes – a robust nonparametric approach to quantify size and overlap of n-dimensional hypervolumes." *Methods in Ecology and Evolution* 7.12, pp. 1503–1513. DOI: 10.1111/2041-210X.12611.

Jørgensen, T. et al. (2012). "A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability." *Molecular Ecology* 21.8, pp. 1989–2003. DOI: 10.1111/j.1365-294X.2011.05287.x.

Kanehisa, M. and P. Bork (2003). "Bioinformatics in the post-sequence era." *Nature Genetics* 33 Suppl, pp. 305–310. DOI: 10.1038/ng1109.

Kant, R., J. Blom, A. Palva, R. Siezen, and W. de Vos (2011). "Comparative genomics of Lactobacillus." *Microbial Biotechnology* 4.3, pp. 323–332. DOI: 10.1111/j.1751-7915.2010.00215.x.

Kao, W.-C., A. H. Chan, and Y. S. Song (2011). "ECHO: a reference-free short-read error correction algorithm." *Genome Research* 21.7, pp. 1181–1192. DOI: 10.1101/gr.111351.110.

Karimzadeh, M. and M. M. Hoffman (2017). "Top considerations for creating bioinformatics software documentation." *Briefings in Bioinformatics*. DOI: 10.1093/bib/bbw134.

Kashtan, N et al. (2014). "Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus." *Science (New York, NY)* 344.6182, pp. 416–420. DOI: 10.1126/science.1248575.

Kattge, J. et al. (2011). "TRY – a global database of plant traits." *Global Change Biology* 17.9, pp. 2905–2935. DOI: 10.1111/j.1365-2486.2011.02451.x.

Katz, D. S. (2017). *FAIR is not fair enough*. Daniel S. Katz's blog. URL: https://danielskatzblog.wordpress.com/2017/06/22/fair-is-not-fair-enough/ (visited on 09/13/2017).

Kaye, J., C. Heeney, N. Hawkins, J. de Vries, and P. Boddington (2009). "Data Sharing in Genomics – Re-shaping Scientific Practice." *Nature reviews. Genetics* 10.5, pp. 331–335. DOI: 10.1038/nrg2573.

Keith, D. A. et al. (2004). "Protocols for listing threatened species can forecast extinction." *Ecology Letters* 7.11, pp. 1101–1108. DOI: 10.1111/j.1461-0248.2004.00663.x.

Keller, A., G. Grimmer, and I. Steffan-Dewenter (2013). "Diverse microbiota identified in whole intact nest chambers of the red mason bee Osmia bicornis (Linnaeus 1758)." *PLoS One* 7829, p. 6.

Keller, A. et al. (2015). "Evaluating Multiplexed next-Generation Sequencing as a Method in Palynology for Mixed Pollen Samples." en. *Plant Biology* 17.2, pp. 558–566. DOI: 10.1111/plb.12251.

Keller, A. et al. (2009). "5.8S-28S rRNA interaction and HMM-based ITS2 annotation." *Gene* 430.1–2, pp. 50–57. DOI: 10.1016/j.gene.2008.10.012.

Keller, A. et al. (2010). "Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees." en. *Biology Direct* 5.1, p. 4. DOI: 10.1186/1745-6150-5-4.

Keller, A., H. Horn, F. Förster, and J. Schultz (2014). "Computational integration of genomic traits into 16S rDNA microbiota sequencing studies." *Gene* 549.1, pp. 186–191. DOI: 10.1016/j.gene.2014.07.066.

Keller, A., G. Grimmer, W. Sickel, and M. J. Ankenbrand (2016). "DNA-Metabarcoding – ein neuer Blick auf organismische Diversität." de. *BIOspektrum* 22.2, pp. 147–150. DOI: 10.1007/s12268-016-0669-0.

Kim, H. T. et al. (2015). "Seven New Complete Plastome Sequences Reveal Rampant Independent Loss of the ndh Gene Family across Orchids and Associated Instability of the Inverted Repeat/Small Single-Copy Region Boundaries." *PLOS ONE* 10.11, e0142215. DOI: 10.1371/journal.pone.0142215.

Klatt, B. K. et al. (2013). "Bee pollination improves crop quality, shelf life and commercial value." *Proceedings of the Royal Society of London B: Biological Sciences* 281.1775. DOI: 10.1098/rspb.2013.2440. eprint: http://rspb.royalsocietypublishing.org/content/281/1775/20132440.full.pdf.

Kleijn, D. and I. Raemakers (2008). "A retrospective analysis of pollen host plant use by stable and declining bumble bee species." *Ecology* 89.7, pp. 1811–1823. DOI: 10.1890/07-1275.1.

Kleyer, M. et al. (2008). "The LEDA Traitbase: a database of life-history traits of the Northwest European flora." *Journal of Ecology* 96.6, pp. 1266–1274. DOI: 10.1111/j.1365-2745.2008.01430.x.

Kodama, Y., M. Shumway, and R. Leinonen (2012). "The sequence read archive: explosive growth of sequencing data." *Nucleic Acids Research* 40 (D1), pp. D54–D56. DOI: 10.1093/nar/gkr854.

Koetschan, C. et al. (2010). "The ITS2 Database III: sequences and structures for phylogeny." *Nucleic Acids Res* 38.Database issue, pp. D275–D279. DOI: 10.1093/nar/gkp966.

Koetschan, C. et al. (2012). "ITS2 Database IV: Interactive taxon sampling for internal transcribed spacer 2 based phylogenies." *Mol Phylogenet Evol* 63.3, pp. 585–588. DOI: 10.1016/j.ympev.2012.01.026.

Kozich, J., S. Westcott, N. Baxter, S. Highlander, and P. Schloss (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform." *Applied and Environmental Microbiology* 79.17, pp. 5112–5120. DOI: 10.1128/AEM.01043-13.

Kraaijeveld, K. et al. (2015). "Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing." *Molecular Ecology Resources* 15.1, pp. 8–16. DOI: 10.1111/1755-0998.12288.

Kress, W. J. and D. L. Erickson (2008). "DNA barcodes: Genes, genomics, and bioinformatics." *Proceedings of the National Academy of Sciences* 105.8, pp. 2761–2762. DOI: 10.1073/pnas.0800476105.

Kress, W. J., K. J. Wurdack, E. A. Zimmer, L. A. Weigt, and D. H. Janzen (2005). "Use of DNA barcodes to identify flowering plants." *Proceedings of the National Academy of Sciences of the United States of America* 102.23, pp. 8369–8374. DOI: 10.1073/pnas.0503123102.

Kress, W. J., C. García-Robledo, M. Uriarte, and D. L. Erickson (2015). "DNA barcodes for ecology, evolution, and conservation." *Trends in Ecology & Evolution* 30.1, pp. 25–35. DOI: 10.1016/j.tree.2014.10.008.

Kripken/Emscripten: Emscripten: An LLVM-to-JavaScript Compiler. https://github.com/kripken/emscripten. Accessed: 2016-09-08.

Krupke, C., G. Hunt, B. Eitzer, G. Andino, and K. Given (2012). "Multiple routes of pesticide exposure for honey bees living near agricultural fields." *PLoS ONE* 7.1. DOI: 10.1371/journal.pone.0029268.

Kuczynski, J. et al. (2011). "Using QIIME to analyze 16S rrna gene sequences from microbial communities." *Current Protocols in Bioinformatics* SUPPL.36. DOI: 10.1002/0471250953.bi1007s36.

Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz (2010). "Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates." *Environmental Microbiology* 12.1, pp. 118–123. DOI: 10.1111/j.1462-2920.2009.02051.x.

Kunnimalaiyaan, M and B. L. Nielsen (1997). "Fine mapping of replication origins (ori A and ori B) in Nicotiana tabacum chloroplast DNA." *Nucleic acids research* 25.18, pp. 3681–3686. DOI: 10.1093/nar/25.18.3681.

Köppler, K., G. Vorwohl, and N. Koeniger (2007). "Comparison of pollen spectra collected by four different subspecies of the honey bee Apis mellifera." *Apidologie* 38.4, pp. 341–353. DOI: 10.1051/apido:2007020.

Lagesen, K. et al. (2007). "RNAmmer: Consistent and rapid annotation of ribosomal RNA genes." *Nucleic Acids Research* 35.9, pp. 3100–3108. DOI: 10.1093/nar/gkm160.

Lan, Y., Q. Wang, J. Cole, and G. Rosen (2012). "Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms." *PLoS ONE* 7.3. DOI: 10.1371/journal.pone.0032491.

Langille, M., W. Hsiao, and F. Brinkman (2008). "Evaluation of genomic island predictors using a comparative genomics approach." *BMC bioinformatics*.

Langille, M. G. I. et al. (2013). "Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences." en. *Nature Biotechnology* 31.9, pp. 814–821. DOI: 10.1038/nbt.2676.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." *Nature Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Laslett, D. and B. Canback (2004). "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." *Nucleic Acids Research* 32.1, pp. 11–16. DOI: 10.1093/nar/gkh152.

Leinonen, R., H. Sugawara, and M. Shumway (2011). "The Sequence Read Archive." *Nucleic Acids Research* 39 (Database issue), pp. D19–D21. DOI: 10.1093/nar/gkq1019.

Letunic, I. and P. Bork (2011). "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." *Nucleic Acids Res* 39.suppl 2, W475–W478. DOI: 10.1093/nar/gkr201. eprint: http://nar.oxfordjournals.org/content/39/suppl_2/W475.full.pdf+html.

Li, B. and C. Dewey (2011). "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome." *BMC Bioinformatics* 12. DOI: 10.1186/1471-2105-12-323.

List, M., P. Ebert, and F. Albrecht (2017). "Ten Simple Rules for Developing Usable Software in Computational Biology." *PLOS Computational Biology* 13.1, e1005265. DOI: 10.1371/journal.pcbi.1005265.

Liu, C. et al. (2012a). "CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences." *BMC genomics* 13, p. 715. DOI: 10.1186/1471-2164-13-715.

Liu, K.-L., A. Porras-Alfaro, C. Kuske, S. Eichorst, and G. Xie (2012b). "Accurate, rapid taxonomic classification of fungal large-subunit rRNA Genes." *Applied and Environmental Microbiology* 78.5, pp. 1523–1533. DOI: 10.1128/AEM.06826-11.

Lohse, M. et al. (2014). "Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data." *Plant, Cell and Environment* 37.5, pp. 1250–1258. DOI: 10.1111/pce.12231.

Lohse, M., O. Drechsel, S. Kahlau, and R. Bock (2013). "OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets." *Nucleic Acids Research* 41 (Web Server issue), W575–581. DOI: 10.1093/nar/gkt289.

Loman, N. and M. Watson (2013). "So you want to be a computational biologist?" *Nature Biotechnology* 31.11, pp. 996–998. DOI: 10.1038/nbt.2740.

Losos, J. B. (2011). "Convergence, Adaptation, and Constraint." *Evolution* 65.7, pp. 1827–1840. DOI: 10.1111/j.1558-5646.2011.01289.x.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology* 15.12. DOI: 10.1186/s13059-014-0550-8.

Ludwig, W., J. Euzeby, and W. Whitman (2010). "Road map of the phyla Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobactere4, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes." *Bergeys Manual of Systematic Bacteriology*, pp. 1–19.

Luo, W., M. Friedman, K. Shedden, K. Hankenson, and P. Woolf (2009). "GAGE: Generally applicable gene set enrichment for pathway analysis." *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-161.

Luque, G. M. et al. (2013). "Ecological effects of environmental change." *Ecology Letters* 16, pp. 1–3. DOI: 10.1111/ele.12050.

Mai, J. C. and A. W. Coleman (1997). "The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants." *Journal of Molecular Evolution* 44.3, pp. 258–271.

Maidak, B. L. et al. (1996). "The Ribosomal Database Project (RDP)." *Nucleic Acids Research* 24.1, pp. 82–85. DOI: 10.1093/nar/24.1.82.

Mailund, T. and C. Pedersen (2004). "QDist - Quartet distance between evolutionary trees." *Bioinformatics* 20.10, pp. 1636–1637. DOI: 10.1093/bioinformatics/bth097.

Mallo, D. and D. Posada (2016). "Multilocus inference of species trees and DNA barcoding." *Phil. Trans. R. Soc. B* 371.1702, p. 20150335. DOI: 10.1098/rstb.2015.0335.

Marchant, R. et al. (2001). "Mid- to Late-Holocene pollen-based biome reconstructions for Colombia." *Quaternary Science Reviews* 20.12, pp. 1289–1308. DOI: 10.1016/S0277-3791(00)00182-7.

Markham, N. R. and M. Zuker (2008). "UNAFold: software for nucleic acid folding and hybridization." eng. *Methods Mol Biol* 453, pp. 3–31. DOI: 10.1007/978-1-60327-429-6_1.

Markowetz, F. (2017). "All biology is computational biology." *PLOS Biology* 15.3, e2002050. DOI: 10.1371/journal.pbio.2002050.

Martin Bland, J. and D. Altman (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." *The Lancet* 327.8476, pp. 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

Martone, M. e. (2014). *Joint Declaration of Data Citation Principles*. FORCE11. San Diego CA. URL: https://force11.org/datacitation (visited on 09/13/2017).

Martín, M. and B. Sabater (2010). "Plastid ndh genes in plant evolution." *Plant Physiology and Biochemistry* 48.8, pp. 636–645. DOI: 10.1016/j.plaphy.2010.04.009.

Marx, V. (2013). "Biology: The big challenges of big data." *Nature* 498.7453, pp. 255–260. DOI: 10.1038/498255a.

Marçais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." *Bioinformatics* 27.6, pp. 764–770. DOI: 10.1093/bioinformatics/btr011.

McDonald, D. et al. (2012). "The Biological Observation Matrix (BIOM) Format or: How I Learned to Stop Worrying and Love the Ome-Ome." *GigaScience* 1, p. 7. DOI: 10.1186/2047-217X-1-7.

McKain, M. R., R. H. Hartsock, M. M. Wohl, and E. A. Kellogg (2017). "Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes." *Bioinformatics* 33.1, pp. 130–132. DOI: 10.1093/bioinformatics/btw583.

McMurdie, P. J. and S. Holmes (2013). "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PLOS ONE* 8.4, e61217. DOI: 10.1371/journal.pone.0061217.

McMurdie, P. J. and S. Holmes (2015). "Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking." *Bioinformatics (Oxford, England)* 31.2, pp. 282–283. DOI: 10.1093/bioinformatics/btu616.

McMurdie, P. J. and J. N. Paulson (2015). *biomformat: An interface package for the BIOM file format*. R/Bioconductor package version 1.0.0.

Merget, B. et al. (2012). "The ITS2 Database." eng. *J Vis Exp* 61. DOI: 10.3791/3806.

Messina, R (2010). *Olea europaea chloroplast, complete genome*. URL: http://www.ncbi.nlm.nih.gov/nuccore/NC_013707.2. 30.06.2015.

Metzker, M. L. (2010). "Sequencing technologies — the next generation." *Nature Reviews Genetics* 11.1, pp. 31–46. DOI: 10.1038/nrg2626.

Meyer, F. et al. (2008). "The Metagenomics RAST Server –a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes." *BMC Bioinformatics* 9, p. 386. DOI: 10.1186/1471-2105-9-386.

Miller, G. (2006). "A Scientist's Nightmare: Software Problem Leads to Five Retractions." *Science* 314.5807, pp. 1856–1857. DOI: 10.1126/science.314.5807.1856.

Miller, J. R., S. Koren, and G. Sutton (2010). "Assembly Algorithms for Next-Generation Sequencing Data." *Genomics* 95.6, pp. 315–327. DOI: 10.1016/j.ygeno.2010.03.001.

Mitchell, R. J., R. E. Irwin, R. J. Flanagan, and J. D. Karron (2009). "Ecology and evolution of plant–pollinator interactions." *Annals of Botany* 103.9, pp. 1355–1363. DOI: 10.1093/aob/mcp122.

Mons, B. et al. (2017). "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud." *Information Services & Use* 37.1, pp. 49–56. DOI: 10.3233/ISU-170824.

Mortazavi, A., B. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature Methods* 5.7, pp. 621–628. DOI: 10.1038/nmeth.1226.

Müller, T., N. Philippi, T. Dandekar, J. Schultz, and M. Wolf (2007). "Distinguishing species." *RNA* 13.9, pp. 1469–1472. DOI: 10.1261/rna.617107.

Mullins, J. and J. Emberlin (1997). "Sampling pollens." *Journal of Aerosol Science*. Sampling and Rapid Assay of Bioaerosols 28.3, pp. 365–370. DOI: 10.1016/S0021-8502(96)00439-9.

Mungall, C. et al. (2007). "A Chado case study: An ontology-based modular schema for representing genome-associated biological information." *Bioinformatics* 23.13, pp. i337–i346. DOI: 10.1093/bioinformatics/btm189.

NCBI Resource Coordinators (2015). "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Res* 43.Database issue, pp. D6–D17.

NCBI Resource Coordinators (2017). "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 45 (D1), pp. D12–D17. DOI: 10.1093/nar/gkw1071.

Nagalakshmi, U. et al. (2008). "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science (New York, N.Y.)* 320.5881, pp. 1344–1349. DOI: 10.1126/science.1158441.

Namouchi, A, X Didelot, U Schöck, and B Gicquel (2012). "After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection." *Genome*.

Neuhaus, I. (2016). *CanvasXpress*.

Nielsen, C. B., M. Cantor, I. Dubchak, D. Gordon, and T. Wang (2010). "Visualizing genomes: techniques and challenges." *Nature Methods* 7.3s, S5–S15. DOI: 10.1038/nmeth.1422.

Nilsson, R. H. et al. (2006). "Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective." *PLOS ONE* 1.1, pp. 1–4. DOI: 10.1371/journal.pone.0000059.

Novichkov, P., I. Ratnere, Y. Wolf, E. Koonin, and I. Dubchak (2009). "ATGC: A database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes." *Nucleic Acids Research* 37.SUPPL. 1, pp. D448–D454. DOI: 10.1093/nar/gkn684.

Nussbaumer, T. et al. (2014). "RNASeqExpressionBrowser-a web interface to browse and visualize high-throughput expression data." *Bioinformatics* 30.17, pp. 2519–2520. DOI: 10.1093/bioinformatics/btu334.

O'Leary, N. A. et al. (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." *Nucleic Acids Research* 44 (D1), pp. D733–D745. DOI: 10.1093/nar/gkv1189.

Odoux, J.-F. et al. (2012). "Territorial biodiversity and consequences on physico-chemical characteristics of pollen collected by honey bee colonies." *Apidologie* 43.5, pp. 561–575. DOI: 10.1007/s13592-012-0125-1.

Ondov, B. D., N. H. Bergman, and A. M. Phillippy (2011). "Interactive metagenomic visualization in a Web browser." *BMC bioinformatics* 12, p. 385. DOI: 10.1186/1471-2105-12-385.

Oteros, J. et al. (2015). "Automatic and Online Pollen Monitoring." *International Archives of Allergy and Immunology* 167.3, pp. 158–166. DOI: 10.1159/000436968.

Pagani, I. et al. (2012). "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata." *Nucleic acids research* 40.Database issue, pp. D571–9. DOI: 10.1093/nar/gkr1100.

Page, R. (2016). "Towards a biodiversity knowledge graph." *Research Ideas and Outcomes* 2, e8767. DOI: 10.3897/rio.2.e8767.

Pang, X., L. Shi, J. Song, X. Chen, and S. Chen (2012). "Use of the potential DNA barcode ITS2 to identify herbal materials." en. *J Nat Med* 67.3, pp. 571–575. DOI: 10.1007/s11418-012-0715-2.

Parameswaran, P. et al. (2007). "A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing." *Nucleic Acids Research* 35.19. DOI: 10.1093/nar/gkm760.

Parducci, L., Y. Suyama, M. Lascoux, and K. Bennett (2005). "Ancient DNA from pollen: A genetic record of population history in Scots pine." *Molecular Ecology* 14.9, pp. 2873–2882. DOI: 10.1111/j.1365-294X.2005.02644.x.

Pareek, C. S., R. Smoczynski, and A. Tretyn (2011). "Sequencing technologies and genome sequencing." *Journal of Applied Genetics* 52.4, pp. 413–435. DOI: 10.1007/s13353-011-0057-x.

Parr, C. S. et al. (2014a). "The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth." *Biodiversity Data Journal* 2, e1079. DOI: 10.3897/BDJ.2.e1079.

Parr, C. S. et al. (2014b). "TraitBank: Practical semantics for organism attribute data." *Semant Web–Interoperability, Usability, Appl an IOS Press J*, pp. 650–1860.

Paten, B. et al. (2011). "Cactus: Algorithms for genome multiple sequence alignment." *Genome Research* 21.9, pp. 1512–1528. DOI: 10.1101/gr.123356.111.

Patro, R., G. Duggal, and C. Kingsford (2015). "Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment." *Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data Using Lightweight-alignment*.

Patro, R., S. M. Mount, and C. Kingsford (2014). "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms." *Nature Biotechnology* 32.5, pp. 462–464. DOI: 10.1038/nbt.2862.

Paulson, J. N., O. C. Stine, H. C. Bravo, and M. Pop (2013a). "Differential abundance analysis for microbial marker-gene surveys." *Nature methods* 10.12, pp. 1200–2. DOI: 10.1038/nmeth.2658.

Paulson, J. N., O. C. Stine, H. C. Bravo, and M. Pop (2013b). "Robust methods for differential abundance analysis in marker gene surveys." *Nature methods* 10.12, pp. 1200–1202. DOI: 10.1038/nmeth.2658.

Picard-Nizou, A. et al. (1995). "Foraging behaviour of honey bees (Apis mellifera L.) on transgenic oilseed rape (Brassica napus L. var. oleifera)." *Transgenic Research* 4.4, pp. 270–276. DOI: 10.1007/BF01969121.

Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." *Trends in genetics : TIG* 24.3, pp. 142–149. DOI: 10.1016/j.tig.2007.12.007.

Praz, C., A. Müller, and S. Dorn (2008). "Host recognition in a pollen-specialist bee: Evidence for a genetic basis." *Apidologie* 39.5, pp. 547–557. DOI: 10.1051/apido:2008034.

Prlić, A. and J. B. Procter (2012). "Ten Simple Rules for the Open Development of Scientific Software." *PLOS Computational Biology* 8.12, e1002802. DOI: 10.1371/journal.pcbi.1002802.

Pruitt, K., T. Tatusova, and D. Maglott (2007). "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Research* 35.SUPPL. 1, pp. D61–D65. DOI: 10.1093/nar/gkl842.

Pull request #67 · PitchInteractiveInc/Phinch. https://github.com/PitchInteractiveInc/Phinch/pull/67. Accessed: 2016-12-22, preview version online at https://blackbird.iimog.org.

Qi, J., B. Wang, and B.-I. Hao (2004). "Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach." *Journal of Molecular Evolution* 58.1, pp. 1–11. DOI: 10.1007/s00239-003-2493-7.

Qiita (2016). https://qiita.ucsd.edu/. Accessed: 2016-09-08.

Queiroz, A. de and J. Gatesy (2007). "The supermatrix approach to systematics." *Trends in Ecology & Evolution* 22.1, pp. 34–41. DOI: 10.1016/j.tree.2006.10.002.

R Core Team (2010). "R: A Language and Environment for Statistical Computing." *R: A Language and Environment for Statistical Computing*.

R Core Team (2013). "R: A language and environment for statistical computing." *R: A Language and Environment for Statistical Computing*.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raes, J. and P. Bork (2008). "Molecular eco-systems biology: towards an understanding of community function." *Nature Reviews. Microbiology* 6.9, pp. 693–699. DOI: 10.1038/nrmicro1935.

Raes, J., I. Letunic, T. Yamada, L. J. Jensen, and P. Bork (2011). "Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data." *Molecular Systems Biology* 7, p. 473. DOI: 10.1038/msb.2011.6.

Rambaut, A. (2017). *FigTree*. URL: http://tree.bio.ed.ac.uk/software/figtree/ (visited on 09/23/2017).

Ranjan, R., A. Rani, A. Metwally, H. S. McGee, and D. L. Perkins (2016). "Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing." *Biochemical and Biophysical Research Communications* 469.4, pp. 967–977. DOI: 10.1016/j.bbrc.2015.12.083.

Rasko, D. A. et al. (2011). "Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany." *The New England journal of medicine* 365.8, pp. 709–717. DOI: 10.1056/NEJMoa1106920.

Ratnasingham, S. and P. D. N. Hebert (2007). "bold: The Barcode of Life Data System (http://www.barcodinglife.org)." en. *Mol Ecol Notes* 7.3, pp. 355–364. DOI: 10.1111/j.1471-8286.2007.01678.x.

Richardson, R. et al. (2015a). "Application of ITS2 Metabarcoding to determine the provenance of pollen collected by honey bees in an Agroecosystem." *Applications in Plant Sciences* 3.1. DOI: 10.3732/apps.1400066.

Richardson, R. T. et al. (2015b). “Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus metabarcoding approach.” *Applications in Plant Sciences* 3.11. DOI: 10.3732/apps.1500043.

Riley, D. R., S. V. Angiuoli, J. Crabtree, J. C. Dunning Hotopp, and H. Tettelin (2012). “Using Sybil for interactive comparative genomics of microbes on the web.” eng. *Bioinformatics* 28.2, pp. 160–166. DOI: 10.1093/bioinformatics/btr652.

Roberts, R., M. Carneiro, and M. Schatz (2013). “The advantages of SMRT sequencing.” *Genome Biology* 14.6. DOI: 10.1186/gb-2013-14-6-405.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics* 26.1, pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

Rougier, N. P. et al. (2017). “Sustainable computational science: the ReScience initiative.” *arXiv:1707.04393 [cs]*. arXiv: 1707.04393.

Roux, S. et al. (2016). “Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses.” *Nature* 537.7622, pp. 689–693. DOI: 10.1038/nature19366.

Ruhl, M. W., M. Wolf, and T. M. Jenkins (2010). “Compensatory base changes illuminate morphologically difficult taxonomy.” *Mol Phylogenet Evol* 54.2, pp. 664–669. DOI: http://dx.doi.org/10.1016/j.ympev.2009.07.036.

Ruoff, K. et al. (2007). “Authentication of the botanical origin of honey using profiles of classical measurands and discriminant analysis.” *Apidologie* 38, pp. 438–452.

Saitou, N. and M. Nei (1987). “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” *Molecular Biology and Evolution* 4.4, pp. 406–425.

Salzberg, S. L. et al. (2011). “GAGE: A critical evaluation of genome assemblies and assembly algorithms.” *Genome Research*, gr.131383.111–. DOI: 10.1101/gr.131383.111.

Sanderson, L.-A. et al. (2013). “Tripal v1.1: A standards-based toolkit for construction of online genetic and genomic databases.” *Database* 2013. DOI: 10.1093/database/bat075.

Sandve, G. K., A. Nekrutenko, J. Taylor, and E. Hovig (2013). “Ten Simple Rules for Reproducible Computational Research.” *PLOS Computational Biology* 9.10, e1003285. DOI: 10.1371/journal.pcbi.1003285.

Sayers, E. et al. (2011). “Database resources of the national center for biotechnology information.” *Nucleic Acids Research* 39.SUPPL. 1, pp. D38–D51. DOI: 10.1093/nar/gkq1172.

Sayers, E. W. et al. (2009). “Database resources of the National Center for Biotechnology Information.” eng. *Nucleic Acids Research* 37.Database issue, pp. D5–15. DOI: 10.1093/nar/gkn741.

Schloss, P. D. et al. (2009). “Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75.23, pp. 7537–7541. DOI: 10.1128/AEM.01541-09.

Schlumbaum, A., M. Tensen, and V. Jaenicke-Després (2008). “Ancient plant DNA in archaeobotany.” *Vegetation History and Archaeobotany* 17.2, pp. 233–244. DOI: 10.1007/s00334-007-0125-7.

Schnell, I., M. Fraser, E. Willerslev, and M. Gilbert (2010). “Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey.” *Arthropod-Plant Interactions* 4.2, pp. 107–116. DOI: 10.1007/s11829-010-9089-0.

Schoch, C. L. et al. (2012). “Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.” *Proc Natl Acad Sci U S A* 109.16, pp. 6241–6246. DOI: 10.1073/pnas.1117018109.

Scholz, M. B., C.-C. Lo, and P. S. Chain (2012). “Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.” *Current Opinion in Biotechnology*. Analytical biotechnology 23.1, pp. 9–15. DOI: 10.1016/j.copbio.2011.11.013.

Schultz, J. and M. Wolf (2009). "ITS2 sequence–structure analysis in phylogenetics: A how-to manual for molecular systematics." *Mol Phylogenet Evol* 52.2, pp. 520–523. DOI: 10.1016/j.ympev.2009.01.008.

Schultz, J., S. Maisel, D. Gerlach, T. Müller, and M. Wolf (2005). "A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota." *RNA* 11.4, pp. 361–364. DOI: 10.1261/rna.7204505.

Schultz, J. et al. (2006). "The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses." en. *Nucleic Acids Res* 34.suppl 2, W704–W707. DOI: 10.1093/nar/gkl129.

Schulz, M., D. Zerbino, M. Vingron, and E. Birney (2012). "Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels." *Bioinformatics* 28.8, pp. 1086–1092. DOI: 10.1093/bioinformatics/bts094.

Schwarz, R. F. et al. (2015). "Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis." *PLOS Med* 12.2, e1001789. DOI: 10.1371/journal.pmed.1001789.

Seibel, P. N., T. Müller, T. Dandekar, J. Schultz, and M. Wolf (2006). "4SALE – A tool for synchronous RNA sequence and secondary structure alignment and editing." en. *BMC Bioinformatics* 7.1, p. 498. DOI: 10.1186/1471-2105-7-498.

Seibel, P. N., T. Müller, T. Dandekar, and M. Wolf (2008). "Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE." en. *BMC Research Notes* 1.1, p. 91. DOI: 10.1186/1756-0500-1-91.

Selig, C., M. Wolf, T. Müller, T. Dandekar, and J. Schultz (2008). "The ITS2 Database II: homology modelling RNA structure for molecular systematics." *Nucleic Acids Res* 36.Database issue, pp. D377–D380. DOI: 10.1093/nar/gkm827.

Shade, A. et al. (2012). "Lake microbial communities are resilient after a whole-ecosystem disturbance." *The ISME journal* 6.12, pp. 2153–2167. DOI: 10.1038/ismej.2012.56.

Shah, N., H. Tang, T. G. Doak, and Y. Ye (2011). "Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 165–176.

Sheffield, C., P. Hebert, P. Kevan, and L. Packer (2009). "DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies." *Molecular Ecology Resources* 9.SUPPL. 1, pp. 196–207. DOI: 10.1111/j.1755-0998.2009.02645.x.

Sickel, W. (2017). "High-throughput biodiversity assessment - Powers and limitations of meta-barcoding, Hochdurchsatzerfassung von Biodiversität - Stärken und Grenzen von Meta-barcoding." Doctoral Thesis. Würzburg: Julius Maximilians Universität.

Sickel, W. et al. (2015). "Increased Efficiency in Identifying Mixed Pollen Samples by Meta-Barcoding with a Dual-Indexing Approach." en. *BMC Ecology* 15.1, p. 20. DOI: 10.1186/s12898-015-0051-y.

Sickel, W. et al. (in press). "Standard method for identification of bee pollen mixtures through meta-barcoding." *COLOSS BEEBOOK* III.

Smit, A., R Hubley, and P Green (2013). *RepeatMasker Open-4.0*. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>. URL: http://www.repeatmasker.org (visited on 09/24/2017).

Smith, A. M. et al. (2017). "Journal of Open Source Software (JOSS): design and first-year review." *arXiv:1707.02264 [cs]*. arXiv: 1707.02264.

Smith, R. et al. (2012). "InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data." *Bioinformatics* 28.23, pp. 3163–3165. DOI: 10.1093/bioinformatics/bts577.

Soergel, D. A. W. (2015). "Rampant software errors may undermine scientific results." *F1000Research* 3. DOI: 10.12688/f1000research.5930.2.

Soininen, E. et al. (2009). "Analysing diet of small herbivores: The efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures." *Frontiers in Zoology* 6.1. DOI: 10.1186/1742-9994-6-16.

Song, J. et al. (2012). “Extensive Pyrosequencing Reveals Frequent Intra-Genomic Variations of Internal Transcribed Spacer Regions of Nuclear Ribosomal DNA.” *PLoS ONE* 7.8. DOI: 10.1371/journal.pone.0043971.

Sowunmi, M. (1976). “The potential value of honey in palaeopalynology and archaeology.” *Review of Palaeobotany and Palynology* 21.2, pp. 171–185. DOI: 10.1016/0034-6667(76)90017-8.

Spooner, D. (2009). “Dna barcoding will frequently fail in complicated groups: An example in wild potatoes.” *American Journal of Botany* 96.6, pp. 1177–1189. DOI: 10.3732/ajb.0800246.

Stajich, J. E. et al. (2002). “The Bioperl toolkit: Perl modules for the life sciences.” eng. *Genome Res* 12.10, pp. 1611–1618. DOI: 10.1101/gr.361602.

Stamatakis, A. (2014). “RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies.” *Bioinformatics* 30.9, pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033.

Stein, L. D. et al. (2002). “The generic genome browser: a building block for a model organism system database.” *Genome Research* 12.10, pp. 1599–1610. DOI: 10.1101/gr.403602.

Stern, D. L. (2013). “The genetic causes of convergent evolution.” *Nature Reviews. Genetics* 14.11, pp. 751–764. DOI: 10.1038/nrg3483.

Stillman, E. and J. Flenley (1996). “The needs and prospects for automation in palynology.” *Quaternary Science Reviews* 15.1, pp. 1–5. DOI: 10.1016/0277-3791(95)00076-3.

Stone, G. and V. French (2003). “Evolution: Have Wings Come, Gone and Come Again?” *Current Biology* 13.11, R436–R438. DOI: 10.1016/S0960-9822(03)00364-6.

Stout, Jane C. and Morales, Carolina L. (2009). “Ecological impacts of invasive alien species on bees.” *Apidologie* 40.3, pp. 388–409. DOI: 10.1051/apido/2009023.

Sugita, S. (1994). “Pollen representation of vegetation in Quaternary sediments: Theory and method in patchy vegetation.” *Journal of Ecology* 82.4, pp. 881–897.

Sullivan, M. J., N. K. Petty, and S. A. Beatson (2011). “Easyfig: a genome comparison visualizer.” eng. *Bioinformatics* 27.7, pp. 1009–1010. DOI: 10.1093/bioinformatics/btr039.

Suzuki, M. and S. Giovannoni (1996). “Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.” *Applied and Environmental Microbiology* 62.2, pp. 625–630.

Söhngen, C. et al. (2016). “BacDive – The Bacterial Diversity Metadatabase in 2016.” *Nucleic Acids Research* 44 (D1), pp. D581–D585. DOI: 10.1093/nar/gkv983.

Talavera, G. and J. Castresana (2007). “Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.” *Systematic Biology* 56.4, pp. 564–577. DOI: 10.1080/10635150701472164.

Taschuk, M. and G. Wilson (2017). “Ten simple rules for making research software more robust.” *PLOS Computational Biology* 13.4, e1005412. DOI: 10.1371/journal.pcbi.1005412.

Taylor, H. and W. Harris (2012). “An emergent science on the brink of irrelevance: A review of the past 8years of DNA barcoding.” *Molecular Ecology Resources* 12.3, pp. 377–388. DOI: 10.1111/j.1755-0998.2012.03119.x.

Touchon, M and E. Rocha (2007). “Causes of insertion sequences abundance in prokaryotic genomes.” *Molecular biology and evolution*.

Trapnell, C. et al. (2010). “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.” *Nature Biotechnology* 28.5, pp. 511–515. DOI: 10.1038/nbt.1621.

TrinotateWeb: Graphical Interface for Navigating Trinotate Annotations and Expression Analyses (2016).

Truong, D. T., A. Tett, E. Pasolli, C. Huttenhower, and N. Segata (2017). “Microbial strain-level population structure and genetic diversity from metagenomes.” *Genome Research* 27.4, pp. 626–638. DOI: 10.1101/gr.216242.116.

Turnbaugh, P. J. et al. (2007). "The human microbiome project." *Nature* 449.7164, pp. 804–10. DOI: 10.1038/nature06244.

Tzedakis, P. (1993). "Long-term tree populations in northwest Greece through multiple Quaternary climatic cycles." *Nature* 364.6436, pp. 437–440.

Uchiyama, I., M. Mihara, H. Nishide, and H. Chiba (2013). "MBGD update 2013: The microbial genome database for exploring the diversity of microbial world." *Nucleic Acids Research* 41.D1, pp. D631–D635. DOI: 10.1093/nar/gks1006.

Underwood, A. P. et al. (2013). "Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak." *Journal of Clinical Microbiology* 51.1, pp. 232–237. DOI: 10.1128/JCM.01696-12.

Uriarte, M. et al. (2010). "Trait similarity, shared ancestry and the structure of neighbourhood interactions in a subtropical wet forest: implications for community assembly." *Ecology Letters* 13.12, pp. 1503–1514. DOI: 10.1111/j.1461-0248.2010.01541.x.

Valentini, A., F. Pompanon, and P. Taberlet (2009a). "DNA barcoding for ecologists." *Trends in Ecology and Evolution* 24.2, pp. 110–117. DOI: 10.1016/j.tree.2008.09.011.

Valentini, A. et al. (2009b). "New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: The trnL approach." *Molecular Ecology Resources* 9.1, pp. 51–60. DOI: 10.1111/j.1755-0998.2008.02352.x.

Valentini, A., C. Miquel, and P. Taberlet (2010). "DNA barcoding for honey biodiversity." *Diversity* 2.4, pp. 610–617. DOI: 10.3390/d2040610.

Vanneste, K., G. Baele, S. Maere, and Y. V. d. Peer (2014). "Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary." en. *Genome Research* 24.8, pp. 1334–1347. DOI: 10.1101/gr.168997.113.

Velasco, R. et al. (2007). "A high quality draft consensus sequence of the genome of a heterozygous grapevine variety." *PLoS ONE* 2.12. DOI: 10.1371/journal.pone.0001326.

Vere, N. de, T. C. G. Rich, S. A. Trinder, and C. Long (2015). "DNA barcoding for plants." *Methods in Molecular Biology (Clifton, N.J.)* 1245, pp. 101–118. DOI: 10.1007/978-1-4939-1966-6_8.

Vincent, J. et al. (2013). "dbWFA: A web-based database for functional annotation of Triticum aestivum transcripts." *Database* 2013. DOI: 10.1093/database/bat014.

Vos, P. et al. (2009). *Bergey's Manual of Systematic Bacteriology*.

Větrovský, T. and P. Baldrian (2013). "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses." *PLoS ONE* 8.2. DOI: 10.1371/journal.pone.0057923.

Wakasugi, T. et al. (1994). "Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii." en. *Proceedings of the National Academy of Sciences* 91.21, pp. 9794–9798.

Wang, Q., G. Garrity, J. Tiedje, and J. Cole (2007). "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." *Applied and Environmental Microbiology* 73.16, pp. 5261–5267. DOI: 10.1128/AEM.00062-07.

Wang, X.-W. et al. (2010). "De novo characterization of a whitefly transcriptome and analysis of its gene expression during development." *BMC Genomics* 11.1. DOI: 10.1186/1471-2164-11-400.

Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10.1, pp. 57–63. DOI: 10.1038/nrg2484.

Wcislo, W. and J. Cane (1996). "Floral resource utilization by solitary bees (Hymenoptera: Apoidea) and exploitation of their stored foods by natural enemies." *Annual Review of Entomology* 41, pp. 257–286.

Wenping, H., Z. Yuan, S. Jie, Z. Lijun, and W. Zhezhi (2011). "De novo transcriptome sequencing in Salvia miltiorrhiza to identify genes involved in the biosynthesis of active ingredients." *Genomics* 98.4, pp. 272–279. DOI: 10.1016/j.ygeno.2011.03.012.

Werner, J. J. et al. (2012). "Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys." *The ISME journal* 6.1, pp. 94–103. DOI: 10.1038/ismej.2011.82.

Westoby, M. and I. J. Wright (2006). "Land-plant ecology on the basis of functional traits." *Trends in Ecology & Evolution* 21.5, pp. 261–268. DOI: 10.1016/j.tree.2006.02.004.

White, T., T. Bruns, S. Lee, and J. Taylor (1990). "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics." *PCR Protocols: A Guide to Methods and Applications*, pp. 315–322.

Whiting, M. F., S. Bradler, and T. Maxwell (2003). "Loss and recovery of wings in stick insects." *Nature* 421.6920, pp. 264–267. DOI: 10.1038/nature01313.

Wicke, S. et al. (2013). "Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape f amily." *The Plant cell* 25.10, pp. 3711–25. DOI: 10.1105/tpc.113.113373.

Wiens, J. J. and M. R. Servedio (2000). "Species delimitation in systematics: inferring diagnostic differences between species." *Proceedings of the Royal Society B: Biological Sciences* 267.1444, pp. 631–636.

Wilcock, C. and R. Neiland (2002). "Pollination failure in plants: Why it happens and when it matters." *Trends in Plant Science* 7.6, pp. 270–277. DOI: 10.1016/S1360-1385(02)02258-6.

Wilkinson, M. D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, sdata201618. DOI: 10.1038/sdata.2016.18.

Williams, N. and C. Kremen (2007). "Resource distributions among habitats determine solitary bee offspring production in a mosaic landscape." *Ecological Applications* 17.3, pp. 910–921. DOI: 10.1890/06-0269.

Wilson, E., C. Sidhu, K. Levan, and D. Holway (2010). "Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis." *Molecular Ecology* 19.21, pp. 4823–4829. DOI: 10.1111/j.1365-294X.2010.04849.x.

Wilson, G. et al. (2014). "Best Practices for Scientific Computing." *PLOS Biology* 12.1, e1001745. DOI: 10.1371/journal.pbio.1001745.

Wilson, G. et al. (2017). "Good enough practices in scientific computing." *PLOS Computational Biology* 13.6, e1005510. DOI: 10.1371/journal.pcbi.1005510.

Winemiller, K. O., D. B. Fitzgerald, L. M. Bower, and E. R. Pianka (2015). "Functional traits, convergent evolution, and periodic tables of niches." *Ecology Letters* 18.8, pp. 737–751. DOI: 10.1111/ele.12462.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." *Proceedings of the National Academy of Sciences of the United States of America* 74.11, pp. 5088–5090.

Woese, C. (1987). "Bacterial evolution." *Microbiological Reviews* 51.2, pp. 221–271.

Wolf, M., J. Friedrich, T. Dandekar, and T. Müller (2005a). "CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures." *In Silico Biology* 5.3, pp. 291–294.

Wolf, M., M. Achtziger, J. Schultz, T. Dandekar, and T. Müller (2005b). "Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures." *RNA* 11.11, pp. 1616–1623. DOI: 10.1261/rna.2144205.

Wolf, M., B. Ruderisch, T. Dandekar, J. Schultz, and T. Müller (2008). "ProfDistS: (profile-) distance based phylogeny on sequence–structure alignments." *Bioinformatics (Oxford, England)* 24.20, pp. 2401–2402. DOI: 10.1093/bioinformatics/btn453.

Wolf, M., S. Chen, J. Song, M. Ankenbrand, and T. Müller (2013). "Compensatory Base Changes in ITS2 Secondary Structures Correlate with the Biological Species Concept Despite Intragenomic Variability in ITS2 Sequences – A Proof of Concept." *PLoS ONE* 8.6, e66726. DOI: 10.1371/journal.pone.0066726.

Wolf, M., C. Koetschan, and T. Müller (2014). "ITS2, 18S, 16S or any other RNA — simply aligning sequences and their individual secondary structures simultaneously by an automatic approach." *Gene* 546.2, pp. 145–149. DOI: 10.1016/j.gene.2014.05.065.

Wolfe, K. H., C. W. Morden, and J. D. Palmer (1992). "Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant." *Proceedings of the National Academy of Sciences of the United States of America* 89.22, pp. 10648–10652. DOI: 10.1073/pnas.89.22.10648.

Woolfe, M. and S. Primrose (2004). "Food forensics: Using DNA technology to combat misdescription and fraud." *Trends in Biotechnology* 22.5, pp. 222–226. DOI: 10.1016/j.tibtech.2004.03.010.

Wu, D., G. Jospin, and J. Eisen (2013). "Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups." *PLoS ONE* 8.10. DOI: 10.1371/journal.pone.0077033.

Wyman, S. K., R. K. Jansen, and J. L. Boore (2004). "Automatic annotation of organellar genomes with DOGMA." *Bioinformatics (Oxford, England)* 20.17, pp. 3252–3255. DOI: 10.1093/bioinformatics/bth352.

Xia, Z. et al. (2011). "RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis." *Plant Molecular Biology* 77.3, pp. 299–308. DOI: 10.1007/s11103-011-9811-z.

Xu, Z., D. Malmer, M. G. I. Langille, S. F. Way, and R. Knight (2014). "Which is more important for classifying microbial communities: who's there or what they can do?" *The ISME journal* 8.12, pp. 2357–2359. DOI: 10.1038/ismej.2014.157.

Yahara, K, X Didelot, M. Ansari, and S. Sheppard (2014). "Efficient inference of recombination hot regions in bacterial genomes." *Molecular biology and*.

Yang, Z. and B. Rannala (1997). "Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method." *Molecular Biology and Evolution* 14.7, pp. 717–724.

Yao, H. et al. (2010). "Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals." *PLoS ONE* 5.10, e13102. DOI: 10.1371/journal.pone.0013102.

Yarza, P. et al. (2008). "The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains." *Systematic and Applied Microbiology* 31.4, pp. 241–250. DOI: 10.1016/j.syapm.2008.07.001.

Yu, D. W. et al. (2012). "Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring." *Methods in Ecology and Evolution* 3.4, pp. 613–623. DOI: 10.1111/j.2041-210X.2012.00198.x.

Zaneveld, J. R. R. et al. (2011). "Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation." *Trends in Microbiology* 19.10, pp. 472–482. DOI: 10.1016/j.tim.2011.07.006.

Zhou, L.-J., K.-Q. Pei, B. Zhou, and K.-P. Ma (2007). "A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil." *American Journal of Botany* 94.3, pp. 477–481. DOI: 10.3732/ajb.94.3.477.

Ziemann, M., Y. Eren, and A. El-Osta (2016). "Gene name errors are widespread in the scientific literature." *Genome Biology* 17, p. 177. DOI: 10.1186/s13059-016-1044-7.

Zuo, G. and B. Hao (2015). "CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy." *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments 13.5, pp. 321–331. DOI: 10.1016/j.gpb.2015.08.004.

biom convert galaxy module. https://toolshed.g2.bx.psu.edu/repository/view_repository?changeset_revision=501c21cce614&id=b3ae8ca9317b000e. Accessed: 2016-12-15.

hdf5 Javascript in a Webbrowser · Issue #29 · HDF-NI/hdf5.node. https://github.com/HDF-NI/hdf5.node/issues/29. Accessed: 2016-09-08.

Coveralls. URL: https://coveralls.io/. 25.06.2015.

Eclipse IDE for JavaScript Web Developers. URL: https://eclipse.org/downloads/packages/eclipse-ide-javascript-web-developers/indigosr2. 30.06.2015.

Eclipse Standard 4.3.2. URL: https://eclipse.org/downloads/packages/eclipse-standard-432/keplersr2. 25.06.2015.

GenBank. URL: //www.ncbi.nlm.nih.gov/genbank/. 30.06.2015.

LASTZ (2015). URL: http://www.bx.psu.edu/~rsharris/lastz/. 30.06.2015.

Thought Works, Continuous Integration. URL: http://www.thoughtworks.com/de/continuous-integration. 30.06.2015.

Travis CI. URL: https://travis-ci.org/. 25.06.2015.

git. URL: https://git-scm.com/. 25.06.2015.

istanbul-js. URL: https://gotwarlost.github.io/istanbul/. 25.06.2015.

textures.js. URL: http://riccardoscalco.github.io/textures/. 25.06.2015.

AFFIDAVIT

I hereby confirm that my thesis entitled *"Squeezing more information out of biological data - development and application of bioinformatic tools for ecology, evolution and genomics"* is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Würzburg, September 2017

Markus Johannes
Ankenbrand

EIDESSTATTLICHE ERKLÄRUNG

Hiermit erkläre ich an Eides statt, die Dissertation *"Mehr aus biologischen Daten herausholen - Entwicklung und Anwendung bioinformatischer Programme für Ökologie, Evolution und Genomik"* eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg, September 2017

Markus Johannes
Ankenbrand

COLOPHON

This thesis was typeset in LaTeX using TeX Live (2015/Debian), `TeXstudio` (v2.10.8), and the typographical `classicthesis` (v4.2) style developed by André Miede.