

Thesis submitted in fulfilment of the
requirements for
Degree of Doctor of Philosophy

Antarctic biodiversity surveys using high
throughput sequencing: understanding
landscape and communities of the Prince
Charles Mountains

Paul Czechowski

December 2015



THE UNIVERSITY
of ADELAIDE

School of Biological Sciences

Difficulties are just things to overcome, after all - Ernest Shackleton

Contents

Abstract	v
Publications, Presentations, Awards	vii
Thesis Declaration	ix
Acknowledgments	x
Chapter Format	xi
1. Antarctic terrestrial biodiversity and environmental metagenetics	3
1.1. Introduction	4
1.2. Technical considerations	6
1.2.1. Extraction of environmental samples	6
1.2.2. High throughput platforms	7
1.2.3. Marker choice	8
1.2.4. Library generation	8
1.2.5. Amplification	9
1.2.6. Sequence analysis	10
1.2.7. Recent improvements of metagenetic HTS approaches	11
1.3. The potential of metagenetics for Antarctic biology	15
1.3.1. Community structures	15
1.3.2. Geographic distribution of Antarctic biota	16
1.3.3. Supporting conservation efforts	17
1.4. Summary and conclusions	17
1.5. Acknowledgements	18
1.6. Authors contributions	18
2. Eukaryotic soil diversity of the Prince Charles Mountains	21
2.1. Introduction	22

2.2.	Methods and materials	25
2.2.1.	Fieldwork, soil storage and DNA extraction	25
2.2.2.	Amplification and library generation	26
2.2.3.	Read processing	26
2.2.4.	Eukaryotic α and β diversity comparison	27
2.2.5.	Distribution of phylotypes across sites	28
2.2.6.	Species-level assignment of phylotypes	30
2.3.	Results	31
2.3.1.	Read processing	31
2.3.2.	Eukaryotic α and β diversity comparison	31
2.3.3.	Distribution of phylotypes across sites	33
2.3.4.	Species-level assignment of phylotypes	34
2.4.	Discussion	35
2.4.1.	Technical considerations	35
2.4.2.	Differences in eukaryotic diversity among three locations	36
2.4.3.	Distribution of highly abundant phylotypes	37
2.4.4.	Validity of species-level taxonomic assignments	38
2.5.	Summary and conclusions	39
2.6.	Acknowledgements	39
2.7.	Supplemental data	40
2.8.	Supplemental information	41
2.8.1.	Methods and Materials	41
2.8.2.	Results and comments	50
3.	Phylotypes and morphotypes of Antarctic invertebrates	59
3.1.	Introduction	60
3.2.	Methods	63
3.2.1.	Samples	63
3.2.2.	DNA extractions	64
3.2.3.	Primers	65
3.2.4.	Amplification and sequencing	66
3.2.5.	Reference data for taxonomic assignments	66
3.2.6.	Generation of phylotype observations	67
3.2.7.	Selection of processing parameters for 18S and COI phylotypes	68
3.2.8.	Concordance between phylotypes and morphotypes	69

3.3.	Results	71
3.3.1.	Selection of analysis parameters	71
3.3.2.	Concordance between morphotypes and phylotypes	71
3.4.	Discussion	72
3.4.1.	Analysis parameters	74
3.4.2.	Detecting cryptic invertebrates	75
3.4.3.	Metagenetic marker choice for Antarctic invertebrates	76
3.5.	Conclusions	77
3.6.	Acknowledgements	78
3.7.	Supporting information	78
3.8.	Supplemental information	79
3.8.1.	Methods and Materials	79
3.8.2.	Tables and Figures	83
3.8.3.	Analysis code	85
4.	Salinity gradients determine invertebrate distribution	90
4.1.	Introduction	91
4.2.	Methods	94
4.2.1.	Fieldwork	94
4.2.2.	Soil geochemical and mineral analysis	94
4.2.3.	Preparation and analysis of environmental observations	97
4.2.4.	Preparation and analysis of biological observations	98
4.2.5.	Constrained ordination	99
4.3.	Results	99
4.3.1.	Environmental data	99
4.3.2.	Biological data	100
4.3.3.	Biological data in relation to environment	100
4.4.	Discussion	101
4.5.	Conclusion	105
4.6.	Data accessibility	106
4.7.	Authors contributions	107
4.8.	Acknowledgements	107
4.9.	Funding	108
4.10.	Supplemental information	109
4.10.1.	Phylotype data generation for 18S and COI	109

4.10.2. Sequence tag selection and amplicon orientations	110
4.10.3. Intermediate results of environmental data processing	111
4.10.4. Intermediate results of biological data processing	111
4.10.5. Intermediate results of biological data in relation to environment	111
4.10.6. Data and analysis scripts, additional figures and tables	112
5. Synthesis	130
5.1. Summary	130
5.1.1. Technical and computational methods	130
5.1.2. Biodiversity information from the Prince Charles Mountains .	130
5.2. High throughput sequencing for Antarctica	132
5.3. Implications and future improvements	135
5.4. Conclusion	137
5.5. Acknowledgements	138
A. Phylotype information chapter 2	139
B. Analysis code chapter 3	162
C. Analysis code chapter 4	190
D. Molecular tagging of amplicons	236
Bibliography	244

Abstract

Antarctic soils are home to small, inconspicuous organisms including bacteria, unicellular eukaryotes, fungi, lichen, cryptogamic plants and invertebrates. Antarctic soil communities are distinct from other soil biota as a consequence of long-term persistence under harsh environmental conditions; furthermore their long history of isolation is responsible for a high degree of endemism. Of major concern is the establishment of non-indigenous species facilitated by human-mediated climate change and increased human activity, threatening the highly specialised endemic species. A lack of baseline information on terrestrial Antarctic biodiversity currently impairs efforts to conserve the unique but still largely unknown Antarctic biota. In this thesis I apply *metagenetic high throughput sequencing* (MHTS) methods to address the deficiency of biological information from remote regions of continental Antarctica, and use the data generated to explore environmental constraints on Antarctic biodiversity.

In Chapter 1, I introduce current issues impeding the generation of baseline Antarctic biodiversity data and evaluate the application of using MHTS techniques. This review highlights the potential of using MHTS approaches using amplicon sequencing to retrieve Eukaryotic biodiversity information from terrestrial Antarctica. In Chapter 2, the eukaryotic diversity of three biologically unsurveyed regions in the *Prince Charles Mountains*, East Antarctica (PCMs) is explored. Total eukaryote biodiversity in the PCMs appears to follow an altitudinal or latitudinal trend, which is less obvious for terrestrial invertebrates. In order to apply MHTS to the study of Antarctic invertebrates, the comparative taxonomic assignment fidelities of metagenetic markers and morphological approaches are explored in Chapter 3. Fidelities of taxonomic assignments to four Antarctic invertebrate phyla differed depending on metagenetic marker, and only application of non-arbitrary sequence processing parameters resulted in these findings. In Chapter 4, I use MHTS-derived biodiversity information to explore the relationship between soil properties and invertebrate biodiversity in the PCMs. Across large spatial scales distribution of phyla Tardigrada and Arachnida and

classes Enoplea (Nematoda) and Bdelloidea (Rotifera) in inland areas are constrained by terrain-age-related accumulation of salts, while other Classes (Chromadorea, Nematoda and Monogonata, Bdelloidea) are better able to tolerate high salinity. In moister, nutrient-rich and more coastal areas, this effect was less pronounced and a higher invertebrate diversity was found.

The methods applied and developed in this thesis are a valuable starting point to advance the collection of biodiversity information across terrestrial Antarctica and other remote habitats. The work presented here provides examples for generation and usage of MHTS information from remote Antarctic habitats, demonstrates how biodiversity information retrieved using different metagenetic markers can be combined, developed methods for assessing the quality of MHTS markers and finally demonstrated the application of MHTS data to investigate the environmental determinants of invertebrate diversity in remote ice-free habitats. Future MHTS biodiversity studies of Antarctic terrestrial habitats should incorporate large sample numbers and use combined data from multiple genetic markers.

Publications, presentations and awards

Publications

- 2013 Laurence J. Clarke, **Paul Czechowski**, Julien Soubrier, Mark I. Stevens, Alan Cooper (2014). Modular tagging of amplicons using a single PCR for high-throughput sequencing. *Molecular Ecology Resources*. 14, 117–121.¹

Conference presentations

- 2014 2014 SCAR Open Science Conference (SCAR, SkyCity, Auckland, New Zealand): **Paul Czechowski**, Duanne White, Laurence J. Clarke, Alan Cooper, Mark I. Stevens (April): High-throughput DNA sequencing reveals Antarctic soil biodiversity.
- 2014 Understanding Biodiversity Dynamics using diverse Data Sources (CBA, Australian National University, Canberra, Australia): **Paul Czechowski**, Duanne White, Laurence J. Clarke, Alan Cooper, Mark I. Stevens (April): High-throughput DNA sequencing reveals Antarctic soil biodiversity.
- 2013 Biodiversity Genomics Conference (CBA, Australian National University, Canberra, Australia): **Paul Czechowski**, Laurence J. Clarke, Alan Cooper, Mark I. Stevens (April): Exploring Distribution and Evolution of Antarctic Invertebrates using High-Throughput Sequencing.

¹See Appendix D

Awards

- | | |
|------|--|
| 2014 | Australian Government's National Taxonomy Research Grant Program - Student Travel Grant, Australian Biological Resources Study, Government of Australia, \$ 1 650- travel support for conference attendance. |
| 2013 | MiSeq Pilot the Possibilities Grant Program, Illumina Australia, One library reagents kit for experimental libraries. |
| 2012 | Small Research Grants Scheme, The Royal Society of South Australia, \$ 1 500,- for laboratory work. |
| 2012 | International Post-Graduate Research Scholarship, The University of Adelaide, \$ 25 849,- p.a. livelihood. |

Thesis declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

A handwritten signature in blue ink, reading "Paul Goodwin". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

Acknowledgments

I thank my supervisor Dr. Mark Stevens of the South Australian Museum for his long dedication towards this project and his efforts in supporting me in this work and in the years before. Mark has always been a kind mentor and supporter, both professionally and personally. I thank Chester Sands of the British Antarctic Survey, and Alejandro Velasco-Castrillon of the University of Adelaide. I would like to thank those working with me at Mount Menzies, in the Mawson Escarpment and Lake Terrasovoje, Fiona Shanhun, Tessa Williams, Josh Scarrow, Adrian Corvino and Nick Morgan. I appreciate comments and advice provided by several other persons, as further listed in individual chapters.

For this project the Australian Antarctic Division provided funding under science project 2355 to Mark Stevens. The Australian Research Council supported this work through funds from linkage grant LP0991985 to Alan Cooper and Mark Stevens. The University of Adelaide supported this project through the International Post-Graduate Research Scholarship to Paul Czechowski.

Chapter format

This thesis is formatted according to guidelines provided by the University of Adelaide². Reference formatting is uniform throughout the thesis. Sectioning of individual chapters follows the guidelines of currently targeted journals:

Chapter 1 *Antarctic Science*

Chapter 2 *Soil Biology & Biochemistry*

Chapter 3 *PLoS ONE*

Chapter 4 *Royal Society Open Science*


As required by the University of Adelaide, Statements of Authorship are preceding individual chapters. Analysis source code is provided in the appendix as indicated within individual chapters, further supplemental material will be available online upon publishing of individual chapters.

²As provided by higher degree research rules for 2015

Statement of Authorship

Title of Paper	Achieving a landscape and community understanding of Antarctic biodiversity: the potential of metagenetic approaches for the retrieval of terrestrial Antarctic survey data
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Publication Style
Publication Details	<p>The ice-free regions of Antarctica are scarce and most of the 0.3% is concentrated along the coastal margins. Human mediated environmental change and introduction of non-indigenous species increasingly threatens biologic communities in these terrestrial habitats. Consequently, the swift retrieval of survey information from Antarctic terrestrial habitats is critical as a baseline to detect community changes and implement conservation measurements in due time, and to conduct biological studies with data mostly unspoiled by human influence. Most studies to date focus on single species and have a limited spatial coverage, and in consequence are limited in detecting large-scale community properties and property changes. Additionally, surveying terrestrial Antarctic communities has been an arduous task due to difficulties in sampling and identification of specimens. The combination of standardized sampling in the field (e.g. from soil, vegetation, or water) combined with metagenetic approaches using high-throughput sequencing (HTS) of environmental DNA is an appealing solution to overcome these limitations. Environmental metagenetic surveys hence have the potential to play an important role for biomonitoring of Antarctic terrestrial habitats on a continental scale.</p>


Principal Author

Name of Principal Author (Candidate)	Paul Czechowski		
Contribution to the Paper	Designed and structured draft manuscripts, wrote manuscript, designed and created figures, collated biologic data from databases.		
Overall percentage (%)	80%		
Signature		Date	9.6.2015


Co-Author Contributions


By signing the Statement of Authorship, each author certifies that:

- the candidate's stated contribution to the publication is accurate (as detailed above);
- permission is granted for the candidate to include the publication in the thesis; and
- the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Stevens		
Contribution to the Paper	Provided advice and contributed ideas for design and structure of draft manuscripts, edited draft and main manuscripts, provided advice and comments on figures.		
Signature		Date	11.6.2015

Name of Co-Author	Laurence Clarke
-------------------	-----------------

Contribution to the Paper	Provided advice and contributed ideas for design and structure of draft manuscripts, edited draft and main manuscripts.		
Signature		Date	15.6.2015

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Edited main manuscript.		
Signature		Date	12.6.2015

Please cut and paste additional co-author panels here as required.

1. Achieving a landscape and community-level understanding of Antarctic terrestrial biodiversity using environmental metagenetics

Paul Czechowski¹, Laurence J. Clarke^{1, 4, 5}, Alan Cooper¹, Mark I. Stevens^{2, 3}

¹ Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005 Australia; ² South Australian Museum, GPO Box 234, Adelaide SA 5000, Australia; ³ School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, SA 5000, Australia; ⁴ Australian Antarctic Division, Channel Highway, Kingston, TAS 7050, Australia; ⁵ Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, TAS 7001, Australia

Abstract Ice-free regions of Antarctica are mostly concentrated along the coastal margins, but scarce throughout the continental interior. Environmental change, including introduction of non-indigenous species, increasingly threaten these unique terrestrial habitats. At the same time, these unique biotic communities subsisting in isolation across the continent are difficult to survey due to logistical constraints, sampling challenges, and problems related to the identification of small and cryptic taxa. However, baseline biodiversity data from remote Antarctic habitats is critical to detect community changes over time, including newly introduced species, in particular with regard to anticipated environmental changes in the southern polar region. The potential of standardised (non-specialist) sampling in the field (e.g. from soil, vegetation, or water) combined with *high-throughput sequencing* (HTS) of metagenetic bulk DNA is an appealing solution to overcome some of these pitfalls. Such metagenetic HTS approaches benefit from being able to process many samples

in parallel, while workflow and data structure can stay highly uniform, and taxonomic identification of small and cryptic organisms can be revised repeatedly, long after data collection. Hence, similar to the rest of the world, metagenetic surveys backed by HTS play an important role for biomonitoring of Antarctic terrestrial habitats across the continent.

Keywords Antarctica, high throughput sequencing, community, terrestrial, gene marker survey, biomonitoring

1.1. Introduction

Although only 0.3% of continental Antarctica is ice-free, Antarctica is home to many organisms including bacteria, unicellular eukaryotes, fungi, lichen, cryptogamic plants and invertebrates that are scattered across the continent, and subsist in isolated, remote, island-like refuges (Convey et al., 2014). Availability of biodiversity information from these Antarctic areas is fundamental for three major reasons. Firstly, such data facilitates the investigation of glacial constraints and effects on current biodiversity (Convey et al., 2009); secondly, it allows investigation of the effects of environmental change on Antarctic ecosystems (Nielsen and Wall, 2013); and lastly, conservation management becomes possible, in light of increasing threats from non-indigenous invasive species (Chown et al., 2012). However, knowledge of terrestrial Antarctic biodiversity is currently limited because the vast majority of Antarctica's ice-free areas remain unstudied (Convey et al., 2014).

Biodiversity research of ice-free refuges in Antarctica is complicated for two reasons: Firstly, logistic difficulties exacerbated by the harsh environmental conditions may limit biological research to the proximity of research stations, when extensive field work is required (Convey, 2010). Secondly, traditional soil biodiversity assessments of many multicellular eukaryotes may include manual sorting and morphological identification of organisms, which are time consuming and require high taxonomic expertise, especially for the cryptic soil life of Antarctica (Velasco-Castrillón et al., 2014). Molecular methods are better suited for the study of Antarctic biota (Rogers, 2007), but may lack resolution when sequence information is not considered (e.g. compared to analysis of Terminal Restriction Fragment Length Polymorphisms – TRFLP's; Makhalanyane et al. 2013) or may be work intensive (e.g. Sanger-sequencing, Fell et al., 2006; Lawley et al., 2004; Velasco-Castrillón et al., 2014).

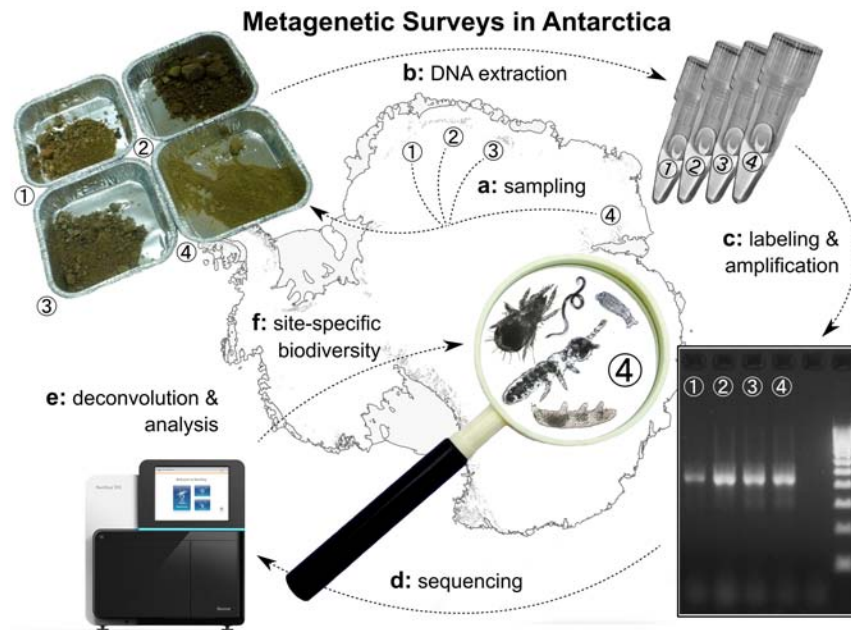


Figure 1.1.: Example workflow for metagenetic analyses of Antarctic environmental samples. A - Sample collection: Samples are collected from a variety of substrates, such as soils at different locations. B - DNA extraction: The entire genetic material is then extracted in bulk from individual samples. C - PCR amplicon library generation: DNA contained in bulk extracts are subsequently amplified with suitable genetic markers and platform-specific sequencing adapters, and multiplex identifier (MID) tags are added. D - Sequencing: The library is processed on a high-throughput-sequencing (HTS) device. E and F - Deconvolution, clustering and / or taxonomy assignment: Reference information is used to assign individual sequences or clusters of similar sequences with taxonomic information after data deconvolution according to sample. F - If phylotypes are shared across several samples, distributional information becomes available. Picture of sequencing device provided courtesy of Illumina Inc. Base layers courtesy of the SCAR Antarctic Digital Database, © 1993-2015 Scientific Committee on Antarctic Research; The National Snow and Ice Data Centre, University of Colorado, Boulder; NASA, Visible Earth Team, <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia 2006.

Readily applied in many other parts of the world (reviewed in Bik et al., 2012; Bohmann et al., 2014), metagenetics present an opportunity to rapidly generate baseline biodiversity information for terrestrial Antarctic habitats (Fig. 1.1; Chown et al., 2015). Metagenetic approaches utilise the genetic material from bulk environmental samples such as soil, water or other substrates (Bohmann et al., 2014). DNA from multiple organisms contained in such samples are then identified either

with traditional Sanger sequencing or more recently using *high throughput sequencing* (HTS) technology (Chown et al., 2015; Cowan et al., 2015). In a global context, HTS-supported metagenetic approaches have been applied to monitoring invasive species and surveying biodiversity over large spatial scales (reviewed in Bohmann et al., 2014). In Antarctica, metagenetic studies, initially based on Sanger sequencing, have enabled the detailed identification of cryptic organisms and communities such as fungi, yeast and invertebrates (Fell et al., 2006; Lawley et al., 2004). More recently, HTS metagenetic studies have been applied to viruses (López-Bueno et al., 2009), bacteria in hypolithic communities, soil and air (Bottos et al., 2014; Makhallanyane et al., 2013), as well as fungal and unicellular eukaryotes of soils (Dreesens et al., 2014; Niederberger et al., 2015).

Collectively, HTS metagenetic approaches provide a promising method to rapidly gather biodiversity information from Antarctic habitats, with the ability to generate large amounts of biodiversity data, from a wide range of taxa, from simple sample collection and uniform laboratory workflows, and data structures. Here, we provide a technical introduction to HTS metagenetic approaches with an Antarctic focus. We then investigate the potential of such metagenetic approaches for Antarctic biodiversity research beyond their current applications.

1.2. Technical considerations

1.2.1. Extraction of environmental samples

As shown in a variety of world-wide studies (reviewed in Bohmann et al. 2014), extraction of DNA suitable for metagenetic analyses is possible from a variety of substrates, which offers unique opportunities to study different environments in Antarctica (Fig. 1.1 a). DNA could be extracted from organisms contained in surface soil (Creer et al., 2010), permafrost (Bellemain et al., 2013), or freshwater benthos of lakes (Hajibabaei et al., 2012). Also, extracts of pre-sorted samples can be analysed (Drummond et al., 2015). When limited starting material is available, preservatives such as ethanol can be used as a DNA source (Shokralla et al., 2010). DNA could also be extracted from faeces (De Barba et al., 2014), such as encountered at the seal and penguin colonies in coastal regions of Antarctica.

Failure to extract DNA representative of the sample (Fig. 1.1 b), so called extraction bias (Pedersen et al., 2014), is a major concern for metagenetic approaches, and

may occur if extraction reagents are unable to penetrate cells of different organisms consistently. Alleviation of extraction bias can be achieved through combination of different extraction methods, and include blending samples prior to extraction, and / or using a large amount of starting material (Delmont et al., 2012, 2013; Taberlet et al., 2012). Different extraction methods or batch-wise application of one extraction method may introduce variable levels of non-template contamination (Salter et al., 2014). Therefore, randomised drawing of sample batches is recommended (Salter et al., 2014). Extraction biases and contamination can be discovered by inclusion of negative and positive controls. Negative controls facilitate the detection of contamination; positive controls of known taxonomic composition are helpful in detecting compositional deviations between the sequence data and sample source (Salter et al., 2014). Consequently, both positive and negative controls help to optimise the DNA extraction process, and may be helpful in streamlining processing parameters in steps following extraction.

1.2.2. High throughput platforms

The recent advance of HTS supported metagenetics can be considered a consequence of continuing development of sequencing platforms by companies such as 454 (Roche, Basel, CH-BS, soon to be discontinued), Illumina (San Diego, US-CA), IonTorrent (Thermo Fisher Scientific, Waltham, US-MA) and others since 2005 (Metzker, 2010). These devices generate substantially larger amounts of sequencing data than chain-termination sequencing (Bohmann et al., 2014), but in comparison produce shorter reads (i.e. ~100–600 base pairs, depending on technology). Using these platforms in conjunction with metagenetic approaches does away with the need of processing mixed DNA templates through clone libraries and hence substantially increases time-efficiency of data generation. The most common approach used to generate metagenetic information with HTS is parallel sequencing PCR-amplified bulk DNA extracts, known as *amplicon sequencing* (Bohmann et al., 2014; Taberlet et al., 2012). Important methodological aspects of amplicon sequencing are described later in this chapter, where I also describe how pitfalls of amplicon sequencing can be alleviated. The platform of choice to conduct metagenetic biodiversity surveys currently appears to be one of the Illumina platforms, due to the large amount of sequences that can be processed, a resulting low sequencing price, and the comparatively low error sequencing error rate (Bokulich et al., 2013; Bragg et al., 2013; Caporaso et al., 2010,

2012). The 454 platform, although soon to be discontinued and often comparatively expensive to use, offers the longest read lengths of all platforms suitable for amplicon sequencing (Van Dijk et al., 2014).

1.2.3. Marker choice

Markers for PCR amplification of mixed DNA templates, such as soil with various organisms or environmental DNA (Fig. 1.1 c) should (a) ideally amplify all taxa with similar efficiency despite potential mismatches between primers and the variety of template molecules (Clarke et al., 2014), (b) amplify target regions short enough to allow amplification of potentially degraded DNA, if environmental DNA is targeted (Bellemain et al., 2013; Coissac et al., 2012; Riaz et al., 2011; Taberlet et al., 2012), (c) exhibit the least possible amount of degenerate bases to allow the application of high annealing temperatures, while decreasing the risk of chimeric amplification (Ahn et al., 2012; Kanagawa, 2003; Lenz and Becker, 2008), and (d) target a gene region for which ample reference data is available, to allow taxonomic identification of phylotypes (see below). Finding a primer pair that possesses these desirable, possibly incompatible, qualities is challenging. Two genes that have been widely applied in single-gene and metagenetic analyses of metazoans are the nuclear 18S ribosomal DNA and mitochondrial *cytochrome c oxidase subunit I* (COI) genes (Wu et al., 2011; Zhan et al., 2014). These markers are favoured due to their long history of application, resulting in comparatively abundant reference data in sequence databases such as GenBank, BOLD and SILVA (Benson et al., 2011; Pruesse et al., 2007; Ratnasingham and Hebert, 2007). However, 18S data may underestimate biodiversity due to low taxonomic resolution, and COI may be impeded by insufficiently conserved primer binding sites across broad taxonomic groups (Deagle et al., 2014; Tang et al., 2012). Similar advantages and disadvantages are found analogously in other marker regions applied in metagenetic studies, for example when targeting fungi using the ITS region, or photosynthetic cryptogams via the *matK* gene (CBOL and Janzen, 2009; Orgiazzi et al., 2013).

1.2.4. Library generation

Preparing DNA for HTS requires addition of platform-specific sequencing adapters, and often sample-specific sequence tags (or *multiplex identifier* – MID - tags) to enable deconvolution of sequence data (Fig. 1.1c,e). Initially, DNA pools were furnished

with sequence tags during polymerase chain reaction (Saiki et al., 1988) via extended primer sequences or via ligation of unmodified primers preceding adapter ligation (Binladen et al., 2007; Meyer et al., 2008). More recently, library generation via long primer sequences carrying both adaptor and sequence tags (*fusion primers*) has become more common (applied in Bik et al., 2012). The application of fusion primers is practical in that it requires a single PCR, but may be costly for large numbers of samples and difficult for primer lengths above ~50 base pairs due to poor PCR performance. In those cases, more work-intensive ligation protocols may be a better choice (Kircher et al., 2012; O'Neill et al., 2013; Stiller et al., 2009). Also of concern is the informed choice of sequence tags. Owing to possible flaws of the underlying algorithms, sequence tags may not meet the intended expectations of robustness towards sequencing errors (Faircloth and Glenn, 2012). It is advisable to only use sequence tags that have been explicitly tested for correct Hamming distances (Hamming, 1950), and hence enable correct deconvolution and error correction (Faircloth and Glenn, 2012).

1.2.5. Amplification

Concordance between the taxonomic composition of a mixed DNA template retrieved from environmental bulk samples and the amplified library requires careful calibration of PCR conditions, i.e. length optimisation of denaturation, annealing and extension steps as well as the correct temperatures for the primer-annealing phase (Fig. 1.1 c). Possible pitfalls include (a) introduction of substitutions and insertion / deletions through polymerase activity (Cline et al., 1996), (b) formation of chimeric molecules in late amplification stages (Kanagawa, 2003), (c) amplification bias when using degenerate primers in combination with high annealing temperatures (Cline et al., 1996; Kanagawa, 2003), and (d) failure to detect rare variants when little replication is applied (Ficetola et al., 2014). Such pitfalls collectively threaten the credibility of the resulting sequence data. They may (a) alter the similarity of phylotypes to reference sequences, (b) result in artificial phylotypes that match several reference sequences, (c) artificially enrich phylotypes whose library molecules matched the PCR primers well, or (d) result in false-negative concealment of phylotypes. Retrieval of higher quality data can be achieved by (a) application of proofreading polymerases (Taberlet et al., 2012), (b) using few and long PCR cycles (Ahn et al., 2012; Kanagawa, 2003; Lenz and Becker, 2008), (c) careful testing of annealing temperature, and

(d) processing three or more PCR replicates (Gilbert et al., 2010). Analogous to the extraction step, positive and negative controls are important to track contamination during the amplification procedure. At the same time positive controls may be a source of cross-contamination and should be treated with care.

1.2.6. Sequence analysis

Foremost, it needs to be noted that processing HTS data (Fig. 1.1 e,f) is not straightforward, and requires a high level of bioinformatics expertise and project specific software selection and fine-tuning at every step. To perform a metagenetic analysis with any given raw data set, firstly an analysis workflow needs to be conceptualised. Then, a variety of software algorithms need to be selected with regard to the analysis steps and study goals; also keeping in mind available computing hardware, methods of library design, employed sequencing technology and data volume (Coissac et al., 2012). Subsequently, it is advisable to test programs individually and in order of application, using small data sets. Here, it may be necessary to generate custom (or at least modify existing) text manipulation scripts, through which data input and output between algorithms is handled.

It is possible that the resulting metagenetic analysis workflow is initiated by marker selection (Riaz et al., 2011), and once sequence data has been generated, several raw data processing steps will follow before the statistical analysis can be attempted (Bik et al., 2012; Bohmann et al., 2014). Raw data preparation typically includes quality filtering, removal of sequence adapters, data deconvolution and chimera removal. The clean raw data then is typically clustered, assigned with taxonomy, and the resulting analysis data is checked for its suitability for the intended statistical analysis.

Although raw data preparation can be achieved with a variety of programs (see Tab. 1.1 for examples), software environments dedicated to metagenetic analysis such as QIIME or MOTHUR (Caporaso et al., 2010; Schloss et al., 2009) frequently offer functionality incorporating whole analysis workflows starting from raw data cleaning, phylotype clustering and basic statistical analyses. These metagenetic software environments themselves generally take advantage of multiple algorithms dedicated to particular sub-routines of analysis workflows. For example, chimera detection may be achieved with UPARSE (Edgar, 2013) in QIIME. Taxonomic assignments, for instance, may be retrieved with BLAST (Altschul et al., 1990) such as in QIIME or MOTHUR. In general software environments such as those above

employ specific sub-algorithms that need to be carefully considered before attempting data preparation and analysis steps.

If dedicated metagenetic analysis environments do not offer desired functionalities for analysis and visualisation of cleaned data, such tasks could be achieved through other available software and possibly linked in via “*glue code*” written in programming languages such as R (R Development Core Team, 2011), Bash or Python (Van Rossum and Drake Jr., 1995). EXPLICIT (Robertson et al., 2013), for example, offers basic visualisation and statistic analysis functionally coupled with a graphical user interface. More powerful, but command-driven, the R environment in particular offers extensive packages for the statistical analysis and visualisation of metagenetic data with packages such as Phyloseq or Vegan (McMurdie and Holmes, 2013; Oksanen et al., 2015; R Development Core Team, 2011).

1.2.7. Recent improvements of metagenetic HTS approaches

For the retrieval of biodiversity information over a large spatial scale (Fig. 1.2) cost-efficient processing of hundreds of samples is desirable and modular application of fewer oligonucleotides decreases cost efficiency, but increases workload, as described below. Tagging individual samples with fusion primers increases the cost of metagenetic HTS studies. Presumably for this reason, numbers of parallel processed samples in many recent global and Antarctic metagenetic studies range from seven to twelve samples (Bik et al., 2012; Dreesens et al., 2014; Niederberger et al., 2015; Roesch et al., 2012). Reducing primer-associated costs is possible through modular combination of multiple sequence tags per sample, thus reducing the amount of unique oligonucleotides required for a metagenetic project. Examples for such modular workflows include using two PCRs to double-tag amplicons for HTS (de Cárcer et al., 2011). Similarly, double-tagging can generate amplicons with minimal work, handling and cost in a single PCR (Clarke et al., 2014). A similar, larger scale modular approach is described elsewhere (Bybee et al., 2011).

PCR biases during library preparation (see above) can be alleviated through the application of hybridisation approaches. In hybridisation approaches, annealing of target DNA to biotinylated oligonucleotide probes are used for library generation (Faircloth and Glenn, 2012; Gnirke et al., 2009; Lemmon and Lemmon, 2012). In comparison to PCR, hybridisation approaches enable retrieval of multiple conserved regions per reaction, perform well in detecting rare DNA and reduce compositional

Table 1.1.: Selection of analysis software for metagenetic data of environmental DNA. Possible tasks related to handling of metagenetic data provided in columns, possible software applications provided in rows. Multi-purpose tool in top rows are suitable for various tasks of sequence analysis. Software environments in the middle section are specialised on metagenetic analysis. Programs and packages in the lowest section are focussed on ecological analysis and / or plotting of results. “X”: Functionality of a given tool for a given task is provided. “(X)” Functionality of a given tool for a given task is provided to some extend but more strongly focussed by other tools. “-”: Functionality is not provided.

Name	Interface	Marker development				Quality check	Trimming	Cleaning	Annotation	Phyloptype analysis	Metadatas analysis	Reference(s)	Web link
AdapterRemoval	Unix shell	-	-	-	X	-	-	-	-	-	-	(Lindgreen, 2012)	https://github.com/slindgreen/AdapterRemoval
Trimmomatic	Unix shell	-	-	-	X	-	-	-	-	-	-	(Lohse et al., 2012)	http://www.usadellab.org/cms/?page=trimmomatic
FastQC	Unix shell, graphical	-	X	-	-	-	-	-	-	-	-	-	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Galaxy	(Graphical interface)	-	X	X	X	X	(X)	(X)	(X)	(X)	-	(Giardine et al., 2005)	https://usegalaxy.org/
Uparse	Unix shell	X	X	X	X	X	(X)	(X)	-	-	-	(Edgar, 2013)	http://www.drive5.com/
OBITools,	Unix shell	X	X	X	X	X	X	(X)	(X)	(X)	(X)	(Boyer et al., 2015)	http://www.grenoble.prabi.fr/trac/OBITools
QIIME	Unix shell	-	X	X	(X)	X	X	(X)	X	X	(X)	(Caporaso et al., 2010)	http://qiime.org/
Mothur	Unix shell	-	X	X	(X)	X	X	(X)	X	X	(X)	(Schloss et al., 2009)	http://www.mothur.org/
Megan	Graphical interface	-	-	-	-	-	X	X	X	X	-	(Huson and Weber, 2013)	http://ab.inf.uni-tuebingen.de/software/megan5/
Phyloseq	R	-	-	-	-	-	-	-	-	X	(X)	(McMurdie and Holmes, 2013)	https://joey711.github.io/phyloseq/
Vegan	R	-	-	-	-	-	-	-	-	(X)	X	(Dixon, 2003; Oksanen et al., 2015)	http://cran.r-project.org/web/packages/vegan/index.html
Explicet	Graphical interface	-	-	-	-	-	-	-	-	X	-	(Robertson et al., 2013)	http://www.explicet.org/

biases in the resulting data without the need for extensive replication (Taberlet et al., 2012). A recent study exemplified sequencing of bacterial DNA derived from environmental samples after enrichment with an hybridisation approach, and demonstrated the described benefits outlined here for mixed-template DNA sources (Denonfoux et al., 2013).

The lengths of genomic regions that can be targeted with single read lengths of a given HTS platform is variable (see section ‘High throughput sequencing platforms’), but usually shorter than the 600–1000 bp that can be achieved from a single read using Sanger sequencing technology. Therefore, recent research has investigated the options of adopting shorter fragments of regions that have been used widely in Sanger sequencing, for example the beginning of the COI gene region or the 18S gene (Leray et al., 2013; Machida and Knowlton, 2012). Other studies have identified new marker regions with short read lengths suitable for HTS technologies, but retain adequate information allowing comparisons with data from the traditional markers (Bellemain et al., 2013; CBOL and Janzen, 2009; Epp et al., 2012). In consequence, now not only bacteria and fungi, but also a growing number various metazoan groups and cryptogams can be targeted with metagenetic HTS approaches. Furthermore, identification of custom marker regions is now possible with bioinformatics tools such as ecoPrimers incorporated into OBI tools (Boyer et al., 2015; Riaz et al., 2011) (see Tab. 1.1). The program ecoPrimers, for example, employs user-curated reference data retrieved from databases such as GenBank (Benson et al., 2011), to identify conserved regions suitable for project specific primer design for mixed template amplification.

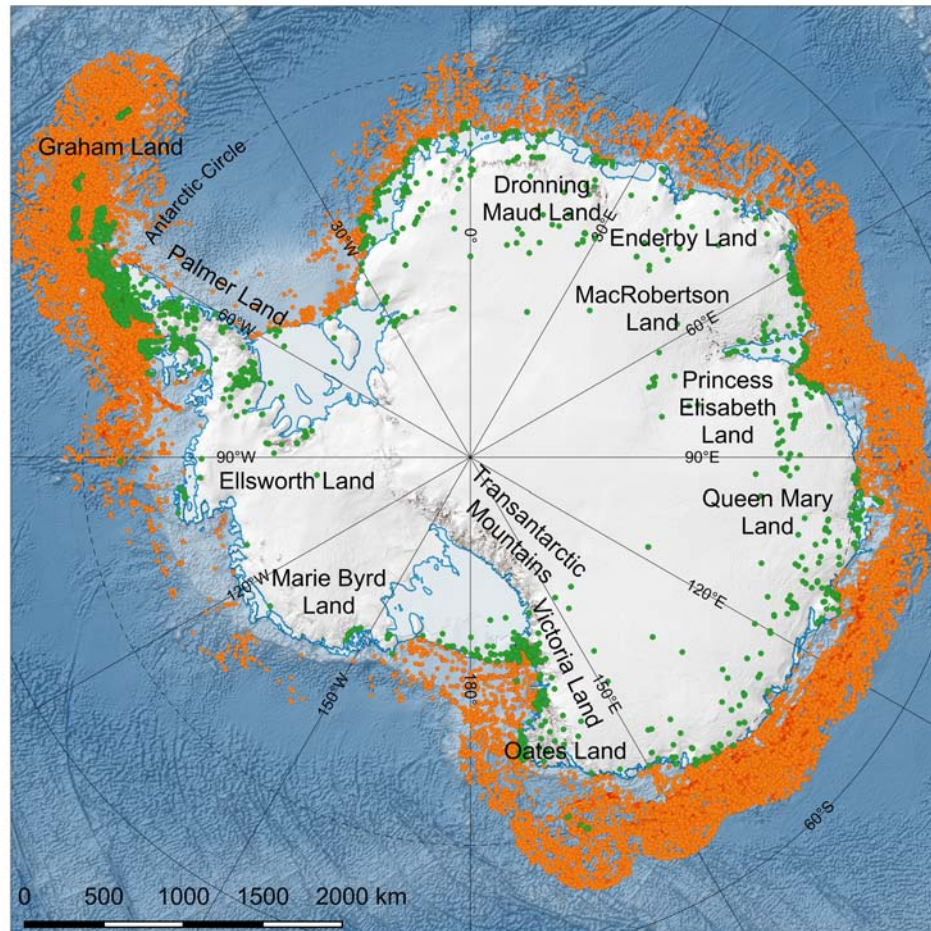


Figure 1.2.: Map of publicly available ground-dwelling Antarctic biodiversity information sourced from GBIF (Flemons et al., 2007). Place names referenced in text. Above-ground observations shown in green, above-sea observations in orange. GBIF download <http://doi.org/10.15468/dl.4b6qco>, 21st of July 2015. Base layers compiled by the Norwegian Polar Institute and distributed in the Quantarctica package. Visit <http://www.quantarctica.org/>. Base layers courtesy of the SCAR Antarctic Digital Database, © 1993-2015 Scientific Committee on Antarctic Research; The National Snow and Ice Data Centre, University of Colorado, Boulder; NASA, Visible Earth Team, <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia 2006. Exposed rock layer from the BEDMAP2 dataset (Fretwell et al., 2011).

1.3. The potential of metagenetics for Antarctic biology

Biodiversity and distribution of soil biota across the ice-free regions of Antarctica is more heterogeneous than anywhere else in the world (Chown et al., 2015; Convey et al., 2014; Ettema and Wardle, 2002). Large distances between habitats, unique geological and glacial histories, different soil compositions and extreme fluctuations of abiotic conditions amplify this heterogeneity (Bintanja et al., 2014; Bockheim, 1997; Marchant and Head, 2007). Densely spaced biological and environmental survey data is required to capture heterogeneous patterns in terrestrial biodiversity and can be provided by metagenetic HTS data in Antarctica, as readily applied in other regions of the world (reviewed in Bik et al., 2012; Bohmann et al., 2014; Chown et al., 2015; Gutt et al., 2012).

1.3.1. Community structures

Community level interaction constitutes an important feature of Antarctic soil biota. Such interactions were believed to be minimal, perhaps owing to the fact that they are hard to measure (Hogg et al., 2006; Velasco-Castrillón et al., 2014). However, biotic community-level interactions are increasingly suggested in facilitating survival in harsh environments, and within biota may be observable through stratified occurrence of different organisms, or the exchange of nutrients between such biotic communities (Colesie et al., 2014; Nakai et al., 2012; Pointing and Belnap, 2012). Community-level organisation has been discovered among Antarctic soil bacteria, soil crusts, lithobiontic communities, eukaryotes in moss pillars and in cyanobacterial mats (Bottos et al., 2014; Buedel and Colesie, 2014; Jungblut et al., 2012; Makhallanyane et al., 2013; Nakai et al., 2012).

Studies describing the community-level organisation of Antarctic terrestrial ecosystems stand to benefit from metagenetic approaches. Possible scopes could include further analysing pro- and eukaryotic diversity in soil crusts and hypolithons, photobiotic and mycobiotic diversity and biogeography of lichen, or the association between fungi and eukaryotes in moss communities, that are still often studied using Sanger sequencing approaches (e.g. Altermann et al., 2014; Fernández-Mendoza et al., 2011; Gokul et al., 2013; Jungblut et al., 2012; Khan et al., 2011; Nakai et al., 2012; Navarro-Noya et al., 2013). At the same time, HTS-based metagenetic

studies are readily seen as a valuable tool to assess the ecological integrity and health of marine environments by providing a harmonised, faster and cheaper means of species identification than morphological approaches (Aylagas et al., 2014). In a similar fashion, HTS-based metagenetic studies may be able to further advance the description of morphologically conserved, rare, or small taxa, as recently demonstrated for Antarctic soil eukaryotes and bacteria, cyanobacterial mats and hypolithic communities (Cowan et al., 2015; Dreesens et al., 2014; Lee et al., 2012; Niederberger et al., 2015). Consequently, for Antarctic biota, HTS-based metagenetic studies similarly offer one simple, cost-efficient workflow, and rich sequence information, that can easily be combined or reanalysed in integrative studies (Chown et al., 2015; Gutt et al., 2012), particularly also including terrestrial microfauna, such as tardigrades, mites, nematodes and springtails (Convey and Stevens, 2007).

1.3.2. Geographic distribution of Antarctic biota

Metagenetic HTS approaches provide opportunities to retrieve baseline biodiversity information from remote regions of Antarctica. Obtaining large-scale survey information of coastal marine fauna may soon be possible using remote imaging approaches (Southwell et al., 2013). For larger patches of lichen and mosses, remote sensing techniques may soon allow rapid retrieval of baseline data across the continent, but may not resolve small patches typically encountered in inland regions (Fretwell et al., 2011). However, obtaining such information for the vast majority of Antarctic biota is more challenging. As a consequence, survey data is sparse for inland invertebrate, cryptogam and fungal communities particularly from large regions across Dronning Maud Land, MacRobertson Land, Oates Land, the Transantarctic Mountains, and Ellsworth Land (Fig. 1.2 and McGaughan et al., 2011; Peat et al., 2007; Velasco-Castrillón et al., 2014). Metagenetic HTS data could potentially provide highly detailed, uniform structured biodiversity information after simple sample collection and facilitate the collection of continent-wide baseline biodiversity information for biota (Bohmann et al., 2014; Chown et al., 2015; Colesie et al., 2014; Gutt et al., 2012). Readily collected sequence data that cannot be assigned with taxonomic reference databases, could be re-examined to retrieve newly recognised taxa at a later stage.

1.3.3. Supporting conservation efforts

Human-mediated environmental changes are anticipated to have profound effects on the spatial extent and structure of Antarctic terrestrial ecosystems (Chown et al., 2015). For example, short-term warming events were shown to cause long-lasting changes in soil moisture, algal chlorophyll a content and shifts in abundance ratios of endemic nematode species in Antarctic terrestrial habitats (Barrett et al., 2008); short-term extremes of high temperatures resulted in shifts of abundance ratios between closely related soil invertebrates with different physiological requirements (Bokhorst et al., 2012). It is anticipated that distribution patterns of Antarctic species will shift southwards and increasingly overlap, possibly eroding the extensive endemism among many Antarctic species (Nielsen and Wall, 2013; Turner et al., 2013).

Elucidating distribution patterns of terrestrial communities and identifying biotic elements most vulnerable to climate change have been deemed some of the most important goals of Antarctic biological conservation (Kennicutt et al., 2015; Sutherland et al., 2009). Antarctica's terrestrial communities are also threatened by non-indigenous species (Frenot et al., 2005; Hughes and Convey, 2014). Such alien introductions may outcompete local endemics in an increasingly accommodating environment (Chown et al., 2012; Hughes and Convey, 2012, 2014). Molecular methods are particularly useful to distinguish Antarctic endemics from non-indigenous species that are not easily detected or are difficult to identify (Hughes and Convey, 2012, 2014). Consequently, the large scale application of HTS metagenetic surveys across the continent would have the potential to generate extensive baseline biodiversity information to support conservation efforts.

1.4. Summary and conclusions

Metagenetic analysis of bulk DNA and environmental DNA is a valuable option to describe the composition and distribution of the cryptic and heterogeneously distributed terrestrial biota of Antarctica. Metagenetic approaches have proven helpful in describing bacterial and hypolithic communities in some ice-free regions of Antarctica, and could similarly be applied to many other taxa on the continent. In comparison to more traditional molecular methods such as TRFLP's or clonal Sanger sequencing, HTS based metagenetic approaches yield large amounts of detailed

data with relatively simple and time-efficient laboratory workflows, coupled with relatively straightforward fieldwork and the ability to process DNA extracted from bulk samples. Multiple laboratory developments have recently improved the cost-efficiency of PCR-based library generation allowing parallel processing of large sample numbers. Drawbacks of amplicon library generation such as the cost of fusion primers, or amplification biases, can be alleviated by alternative library preparation methods. HTS-based metagenetic studies will be a useful tool to assess the ecological integrity and health of Antarctic habitats by providing a consistent and efficient means of species identification. When applied to large sample numbers, across a large spatial scale and multiple biota, HTS based metagenetic approaches will allow integration and understanding of Antarctic terrestrial biodiversity on a continental scale.

1.5. Acknowledgements

The Australian Antarctic Division provided funding under science project 2355 to M.S. The Australian Research Council supported this work through funds from linkage grant LP0991985 to A.C. and M.S. The University of Adelaide supported this project through an International Post-Graduate Research Scholarship to P.C.


1.6. Authors contributions

P.C. prepared, edited and revised the manuscript; M.S. and L.C edited and revised the manuscript.

Statement of Authorship

Title of Paper	A new era of Antarctic exploration: using high-throughput sequencing to reveal eukaryotic soil diversity		
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Publication Style		
Publication Details	<p>Studies of Antarctic eukaryotes have been hampered by their morphological conservatism, small size and the logistical constraints of remote field work, resulting in a deficiency of baseline biodiversity information about Antarctic terrestrial habitats. The application of High-Throughput Sequencing (HTS) in metagenetic approaches is a promising alternative for eukaryotic biodiversity surveys. To test the feasibility of an HTS-based metagenetic approach to examine eukaryotic biodiversity information from remote terrestrial Antarctic habitats, we sequenced 18S rDNA amplicons of twelve Antarctic bulk-soil DNA extracts retrieved from three sampling regions (four samples per region) with the Illumina MiSeq platform. After isolating eukaryotic phylotypes and an initial network visualization, we used rarefied data to compare four α diversity metrics between the three regions. Weighted and unweighted inter-sample UniFrac distances were used for β-diversity comparisons among the twelve individual samples. Furthermore, we analysed the distribution among the most abundant phylotypes and phylotype groups across individual samples. Lastly, we checked the validity of species-level taxonomic assignments. We retrieved 8,126 phylotypes (100%) of which (29.5%) were eukaryotic. Phylotype numbers were lowest for region Mount Menzies (73°S; 3,330 m), intermediate at region Mawson Escarpment (73°S; 807 m,) and highest at region Lake Terrasovoje (70°S; 173 m). Network visualization of unrarefied data and individual samples performed well in comparing diversity and richness, but analysis of rarefied data, despite being combined by sampling region appropriate PCR conditions, was impaired by low read coverage. PCoA of unweighted and weighted UniFrac distances between individual samples reflected patterns observed in network visualisation. Taxonomic information assigned to highest-abundant phylotypes across all individual samples was concordant with expected taxonomic composition of the sampling regions and the most widespread phylotypes were fungal, followed by non-algal protists. In all sampling regions species-level assignments included taxa that are likely to occur in Antarctica. The relation between unrarefied phylotype observations among individual samples indicates increasing eukaryotic richness and diversity with decreasing latitude and altitude. The application of the UniFrac measure could further detail altitude-related richness and diversity changes across individual samples and sampling regions. Approximate species-level taxonomy assignment is possible for Antarctic terrestrial eukaryotes, but needs to be evaluated carefully due to current limitations of reference databases. We show that application of HTS can provide a rapid survey of eukaryotic diversity in Antarctica</p>		

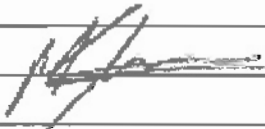
Principal Author


Name of Principal Author (Candidate)	Paul Czechowski		
Contribution to the Paper	Participated in field work, handled samples during extraction and sub-sampling. Designed and applied experiments and analysis approaches, interpreted results, designed and structured draft manuscripts, wrote manuscript, designed and created figures and tables.		
Overall percentage (%)	80%		
Signature		Date	9.6.2015

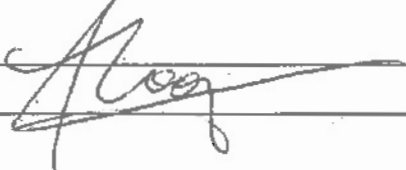
Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Stevens		
Contribution to the Paper	Retrieved funding for, organized, planned, coordinated field campaign in Antarctica. Contributed ideas on experiment and analysis design, interpretation of results, and manuscript structure, revised and edited draft and main manuscripts, contributed ideas for figures and tables. Facilitated access to laboratory facilities, equipment and reagents necessary for soil analysis.		
Signature		Date	11.6.2015

Name of Co-Author	Laurence Clarke		
Contribution to the Paper	Contributed ideas on experiment and analysis design, interpretation of results, and manuscript structure, revised and edited draft and main manuscripts, contributed ideas for figures and tables.		
Signature		Date	15.6.2015

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Contributed ideas on experiment and analysis design. Provided access to laboratory facilities, equipment and reagents necessary for molecular analysis. Provided and facilitated access to computational infrastructure for data analysis. Edited main manuscript.		
Signature		Date	12.6.2015

Please cut and paste additional co-author panels here as required.

2. Antarctic eukaryotic soil diversity of the Prince Charles Mountains revealed by high-throughput sequencing

Paul Czechowski¹, Laurence J. Clarke^{1, 4, 5}, Alan Cooper¹, Mark I. Stevens^{2, 3}

¹ Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005 Australia; ² South Australian Museum, GPO Box 234, Adelaide SA 5000, Australia; ³ School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, SA 5000, Australia; ⁴ Australian Antarctic Division, Channel Highway, Kingston, TAS 7050, Australia; ⁵ Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, TAS 7001, Australia

Abstract Studies of Antarctic eukaryotes have been hampered by their morphological conservatism, small size and the logistical constraints of remote field work, resulting in a deficiency of baseline biodiversity information about Antarctic terrestrial environments. The application of high throughput sequencing (HTS) in metagenetic approaches is a promising alternative. Here, we apply HTS approaches to the hitherto largely unsurveyed micro-eukaryote fauna of the Prince Charles Mountains, East Antarctica. We sequenced 18S rDNA amplicons of twelve Antarctic bulk-soil DNA extracts, retrieved from three sampling regions (four bulk-soil extracts per sampling region). After isolating eukaryotic phylotypes with a stringent filtering approach and initial network visualisation, we firstly used rarefied data to compare four α diversity metrics between the three regions. Weighted and unweighted inter-sample UniFrac distances were then used for β diversity comparisons among rarefied data. Furthermore, we analysed the distribution of the most abundant phylotypes

and phylotype groups. Lastly, we checked the validity of species-level taxonomic assignments using two different taxonomy assignment approaches. Phylotype numbers in un-rarefied data compared across regions were lowest for Mount Menzies (73°S; 3,330 m), intermediate at Mawson Escarpment (73°S; 807 m), and highest at Lake Terrasovoje (70°S; 173 m), likely due to low biological load at the higher latitude and elevation inland sites. Analysis of rarefied data indicated differences in Shannon diversity between Mawson Escarpment and Lake Terrasovoje. PCoA of weighted UniFrac distances between samples from Mawson Escarpment and Lake Terrasovoje indicated changes in community composition in relation to elevation of the sampling locations. The most widespread phylotypes were fungal, followed by non-algal protists. Species-level assignments included known Antarctic taxa in all sampling regions. We show that HTS can provide a rapid survey of the micro-eukaryote fauna to provide baseline biodiversity information for remote environments in the Prince Charles Mountains.

Keywords Antarctic eukaryotes, environmental DNA, high throughput sequencing (HTS), biodiversity survey, Antarctica

2.1. Introduction

Biodiversity of remote Antarctic habitats is key to understanding the history of the Antarctic continent, the biological effects of climate change, as well as for conservation efforts, but many regions of Antarctica remain unsurveyed (Kennicutt et al., 2015; McGeoch et al., 2015). Antarctic soils are home to organisms including bacteria, unicellular eukaryotes, fungi, lichen, cryptogamic plants and invertebrates (Convey et al., 2014). These soil communities are distinct from other soil biota as a consequence of long-term persistence under harsh environmental conditions; furthermore their long history of isolation is responsible for a high degree of endemism (Convey et al., 2008). Simplicity and endemism make Antarctic soil communities interesting for a variety of ecological questions e.g.: Changes in biodiversity patterns in simple communities such as increasing population densities of mites (Kennedy, 1994) and nematodes (Convey, 2003) can be important indicators of human impact and environmental change in terrestrial Antarctica (Nielsen and Wall, 2013). Identifying such indicators in terrestrial Antarctica could also help to understand human-mediated biodiversity changes in more complex temperate and tropical ecosystems and their effect on ecosystem processes, such as decomposition and energy flow (Wall and Virginia,

1999). Studies on the endemism of Antarctic biota revealed that many terrestrial habitats might have become available for re-colonisation since the beginning of the current inter-glacial period (17 000 years ago) (Magalhaes et al., 2012; Stevens and Hogg, 2003). However, there is also evidence that some regions remained ice-free and inhabited for much longer (Convey and Stevens, 2007). Today, human influence increasingly threatens the unique Antarctic soil communities through human-mediated climate change, increasing risk of pollution, and the introduction of non-indigenous organisms which may outcompete endemics in an increasingly accommodating environment. Successful conservation of Antarctic environments in the face of these threats requires biodiversity information (Chown et al., 2012; Terauds et al., 2012; Turner et al., 2013). Unfortunately, this information is missing for many remote ice-free areas of continental Antarctica, such as Dronning Maud Land, large regions of the Transantarctic Mountains (McGaughan et al., 2011), and the Prince Charles Mountains (Terauds et al., 2012).

Biodiversity research in Antarctica is complicated for two main reasons: Firstly, logistic difficulties exacerbated by the harsh environmental conditions typically limits biological research in Antarctica to the proximity of stations when extensive field work is required (Convey, 2010). Secondly, traditional soil biodiversity assessments including manual sorting and morphological identification of organisms are time consuming and require taxonomic expertise, especially for the cryptic soil fauna of Antarctica (Velasco-Castrillón et al., 2014). Molecular methods are better suited for the study of Antarctic biota (Rogers, 2007), but may lack resolution when sequence information is not considered (e.g. in analysis of Terminal Restriction Fragment Length Polymorphisms – TRFLPs; Dreesens et al., 2014; Makhanyane et al., 2013) or may be work intensive (e.g. Sanger-sequencing; Fell et al. 2006; Lawley et al. 2004; Velasco-Castrillón et al. 2014).

High-Throughput Sequencing (HTS) of environmental samples constitutes an interesting opportunity to generate biodiversity information from remote Antarctic habitats (Chown et al., 2015), as it is faster than clonal Sanger sequencing, and field work is simple in comparison to traditional morphological surveys (reviewed in Bik et al., 2012; Bohmann et al., 2014). Hence, HTS based metagenetic studies have been used to monitor invasive species, survey biodiversity over large spatial scales, and provide valuable snapshots of biodiversity for future conservation efforts (Bohmann et al., 2014; Chown et al., 2015; Gutt et al., 2012). Metagenetics in Antarctica have been used to examine viruses (López-Bueno et al., 2009), bacteria in hypolithic

communities (Makhalanyane et al., 2013), soil (Teixeira et al., 2010), air (Bottos et al., 2014), as well as fungi and other eukaryotes (Dreesens et al., 2014; Niederberger et al., 2015; Pointing et al., 2009; Rao et al., 2012).

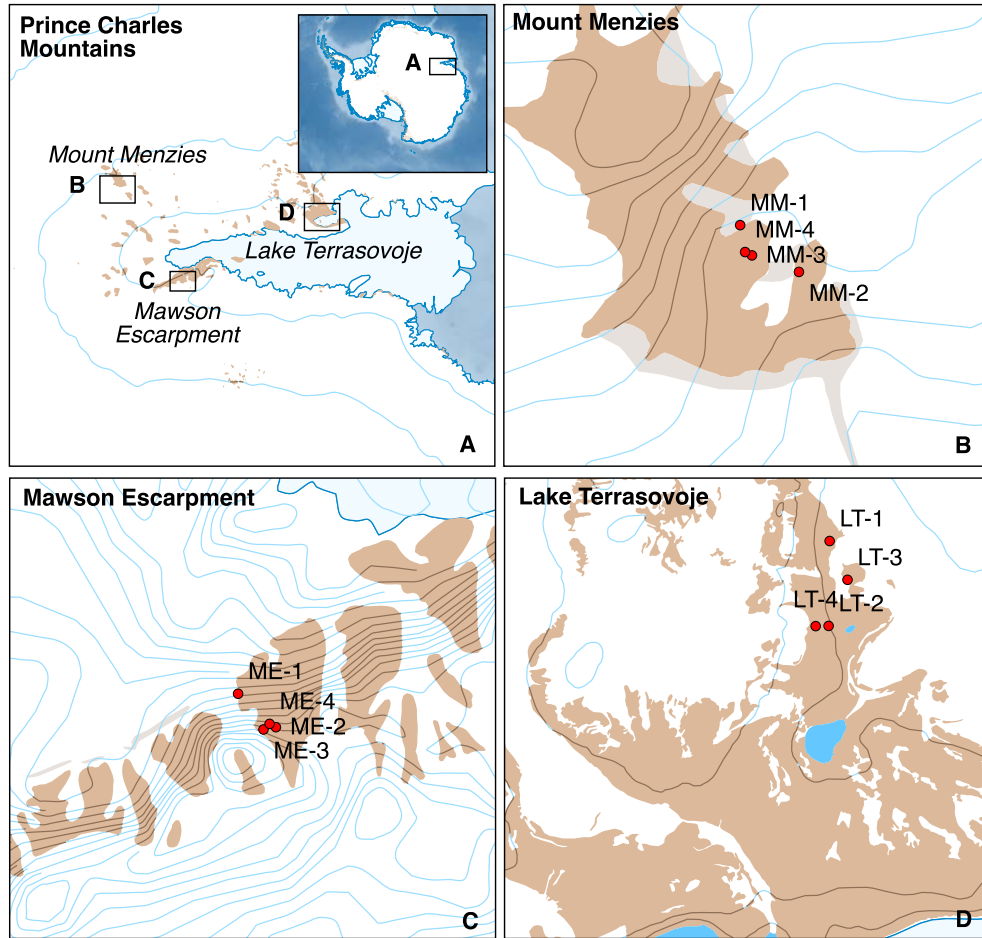


Figure 2.1.: Map of the sampling area in the Prince Charles Mountain, East Antarctica. Sample names correspond to library identifiers used in laboratory work and analysis. Brown: ice-free areas, white: glaciated area, light blue: fresh water lakes or ice fields, dark blue: sea ice or seawater. Base layers compiled by the Norwegian Polar Institute and distributed in the Quantarctica package. Visit <http://www.quantarctica.org/>. Base layers courtesy of the SCAR Antarctic Digital Database, © 1993-2015 Scientific Committee on Antarctic Research; The National Snow and Ice Data Centre, University of Colorado, Boulder; NASA, Visible Earth Team, <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia 2006.

Here we apply a metagenetic HTS approach to explore the micro-eukaryotic soil

biodiversity in three ice-free regions of the Prince Charles Mountains (PCMs) in Eastern Antarctica (Fig. 2.1). With few exceptions (Cremer et al., 2004; Skotnicki et al., 2012; Wagner et al., 2004), the PCMs remain biologically unsurveyed, hindering conservation planning (Terauds et al., 2012). Amplicons of nuclear 18S ribosomal DNA (18S) were generated from bulk soil extracts and sequenced using the Illumina MiSeq platform. Using these data we aimed to: (i) determine any differences in eukaryotic diversity among the sampling regions, (ii) determine whether highly abundant phylotypes in individual samples are widespread or restricted to particular regions, and (iii) examine the validity of species-level taxonomic assignments of Antarctic phylotypes using two different taxonomy assignment approaches.

2.2. Methods and materials

2.2.1. Fieldwork, soil storage and DNA extraction

Fieldwork was conducted during the austral summer of November 2011 / January 2012 at Mount Menzies, Mawson Escarpment and Lake Terrasovoje (Fig. 2.1). Satellite imagery was used to determine several soil-sampling locations within each of the three regions based on broader glaciological and geological properties (bedrock, moraine lines and altitude). Within each region, four sites were then opportunistically chosen for sampling, twelve samples in total. At each sampling site a maximum of 500 g soil was collected from the top 10 cm of the substratum by combining five subsamples from the corners and centre of a one metre square quadrat into a sterile WhirlPak bag (Nasco, Fort Atkinson, US-WI; protocol after Magalhaes et al. 2012). Sample contamination was minimised by wearing nitrile gloves and cleaning equipment with wipes soaked in 70 % ethanol. In the field, samples were stored at -30 to +4 °C in insulated containers (Coleman, Wichita, US-KS). Samples were transported and stored at -20 °C.

DNA extraction was performed at the South Australian Research and Development Institute (SARDI) using a method optimised for the retrieval of DNA from different soil types and the retrieval of invertebrates in agricultural ecosystems for plant pathogen detection (Haling et al., 2011; Huang et al., 2013; Ophel-Keller et al., 2008; Pankhurst et al., 1996), that processes 400 g of starting material. Cross contamination during extraction was detected by measuring the concentration of blank extractions. DNA was stored at -20 °C (SARDI) and at -60 °C (University of Adelaide).

2.2.2. Amplification and library generation

PCR and sequencing primer sequences were sourced from the 18S rRNA amplification protocol 4.13 of the Earth Microbiome Project, as well as groups specialising in developing HTS methods (Gilbert et al., 2010; Parfrey et al., 2014). Fusion primers were designed for use with the Illumina platform (project specific design detailed in supplemental information). Twofold PCR replication was chosen to evaluate the feasibility of amplifying large numbers of samples in subsequent projects. Amplifications were carried out in a volume of 20 μ l, with 2 μ l of template, 1.5 mM $MgCl_2$, 1x AmpliTaq Gold buffer (Thermo Fisher Scientific, Waltham, US-MA), 0.25 mM of each dNTP, 0.5 μ M of forward and reverse primer and 1.25 units AmpliTaq Gold (Thermo Fisher Scientific, Waltham, US-MA). After initial denaturation at 94 °C for 3 min, PCR was performed with 35 cycles of 94 °C for 45 s, 57 °C for 1 min, and 72 °C for 1:30 min, with final elongation of 10 min at 72 °C. To monitor and remove contamination, no-template controls were included in the amplification, sequencing and analysis procedure. Amplicons were visualised on 2 % agarose gels, then duplicates were pooled and purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, US-CA). Amplicon DNA concentrations were quantified using a Qubit 2.0 fluorimeter (dsDNA HS Assay - Thermo Fisher Scientific, Waltham, US-MA) and a 2200 TapeStation with High Sensitivity D1K ScreenTapes (Agilent Technologies, Santa Clara, US-CA). Amplicons were then combined in equimolar ratios (4.5 pmol, see supplemental information) and sequenced in both directions on an Illumina MiSeq using reagents kit v2, 300 cycles (Illumina, San Diego, US-CA; 150 bp paired-end reads). The MiSeq run was shared with other, unrelated projects.

2.2.3. Read processing

Sequencing adapters and primers were removed using Trimmomatic v0.30 (Lohse et al., 2012) and AdapterRemoval v1.1 (Lindgreen, 2012). AdapterRemoval was used to merge paired-end reads. Quality filtering was performed by each of the two programs; the final merged reads were filtered a third time using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/; accessed 1.1.2014, see supplemental information for detailed settings) to further increase quality of remaining reads. Subsequent steps were performed using QIIME v1.8 (Caporaso et al., 2010). Putative chimeric sequences were removed with USEARCH v5.2.236 (Edgar, 2010, 2013) by comparison to the SILVA database v108, which provides curated reference data for 18S

rDNA amplicons across a wide range of eukaryotes (Pruesse et al., 2007). Sequences were clustered *de novo*, as comprehensive reference data is not available for Antarctic soil biota, the default clustering threshold of 97 % similarity was used similar to other Antarctic 18S studies (Lawley et al., 2004). In a first taxonomy assignment (using QIIME), each phylotype (i.e. sequence cluster) was assigned taxonomy from SILVA database v111 (ftp://thebeast.colorado.edu/pub/QIIME_nonstandard_referencedb/; accessed 30.6.2014) using UCLUST v1.2.22q (Edgar, 2010, 2013) and the default 90 % similarity threshold of QIIME. Detailed information is provided in the supplemental information.

Phylotypes were filtered stepwise using QIIME. Firstly, phylotypes without taxonomic assignments were removed, resulting in retention only of known eukaryotic phylotypes (i.e. phylotypes contained in the reference database and assigned to reference information with 97 % similarity). Subsequently, phylotypes present in PCR controls were subtracted from the Antarctic sample data, preventing the analysis of reagent contamination. Lastly, phylotypes not assigned to eukaryotes were removed. Sequence and unique phylotype counts for samples after each filtering step are listed in supplemental Tab. 2.4. An initial network visualisation of phylotype distribution across sampling regions was used for further quality evaluation of un-rarefied eukaryotic phylotype data.

2.2.4. Eukaryotic α and β diversity comparison

Eukaryotic α and β diversity were estimated using rarefied data with QIIME and R v3.1.1 (R Development Core Team, 2011). Four α diversity metrics were estimated and β diversity was analysed through *Principal Coordinate Analysis* (PCoA) of abundance-weighted and unweighted UniFrac distances (Lozupone and Knight, 2005).

Minimum sequence count for samples from Mount Menzies and Mawson Escarpment was 163 and 460, respectively (see supplemental Tab. 2.4 and Fig. 2.6) Consequently, two rarefaction depths were used for α and β diversity comparison of known eukaryotes between regions, 160 for the inclusion of all samples, 460 for comparison of Mawson Escarpment and Lake Terrasovoje only. Sample data were rarefied in steps of ten, with ten iterations per step for error estimation. Good's index was used to estimate coverage (i.e. percentage of the total sequences represented in rarefied data; Good, 1953). As a second metric, the number of observed known phylotypes was

included. The network showed both highly and less abundant phylotypes (Fig. 2.2), hence Shannon diversity and Simpson's index were included as α diversity metrics; Shannon's diversity to increase the weighting of rare phylotypes and Simpson's index to increase the weighting of more abundant phylotypes (DeJong, 1975; Shannon, 1948; Simpson, 1949). Comparison of α diversity metrics between regions was initiated using rarefaction plots (supplemental Fig. 2.6). Analysis of variance (ANOVA) and Tukey's *honest significant difference* tests (HSD) (Tukey, 1949) were then used to test for significant differences at both rarefaction depths.

To detect location specific similarities between individual samples among known eukaryotes in correlation to latitude and altitude as proxy variables, inter-sample UniFrac distances were used in the PCoA. These UniFrac distances were calculated using a phylogenetic tree generated by FASTTREE (Price et al., 2009). For this tree an alignment was generated from sequences representing each known eukaryotic phylotype found across all twelve samples using PYNAST (Caporaso et al., 2010), this alignment was filtered to exclude 10% of the most entropic positions as suggested by the QIIME documentation. Error estimation was conducted through repeated distance matrix calculation in conjunction with Jackknife resampling (Efron and Stein, 1981), sampling depth was set to 160 (all regions) and to 460, respectively (Mawson Escarpment and Lake Terrasovoje). Master trees for Jackknife resampling were calculated using un-subsampled data.

2.2.5. Distribution of phylotypes across sites

We explored whether highly abundant known eukaryotic phylotypes have restricted or widespread distributions in the Prince Charles Mountain using a combination of QIIME functions and Bash scripts, analogous to the approach of Lawley et al. (2004). Eukaryotic phylotype data was collated so that it contained only phylotypes shared across 12 to one samples using QIIMEs sub-setting functions. To prevent spurious results only phylotypes contributing at least 1% to total sample abundance were included. Identification of highly abundant phylotypes in resulting soil sample groups were determined with non-parametric ANOVA (Kruskal and Wallis, 1952) as implemented in QIIME. All phylotypes were then grouped in categories and mapped as in Lawley et al. (2004).

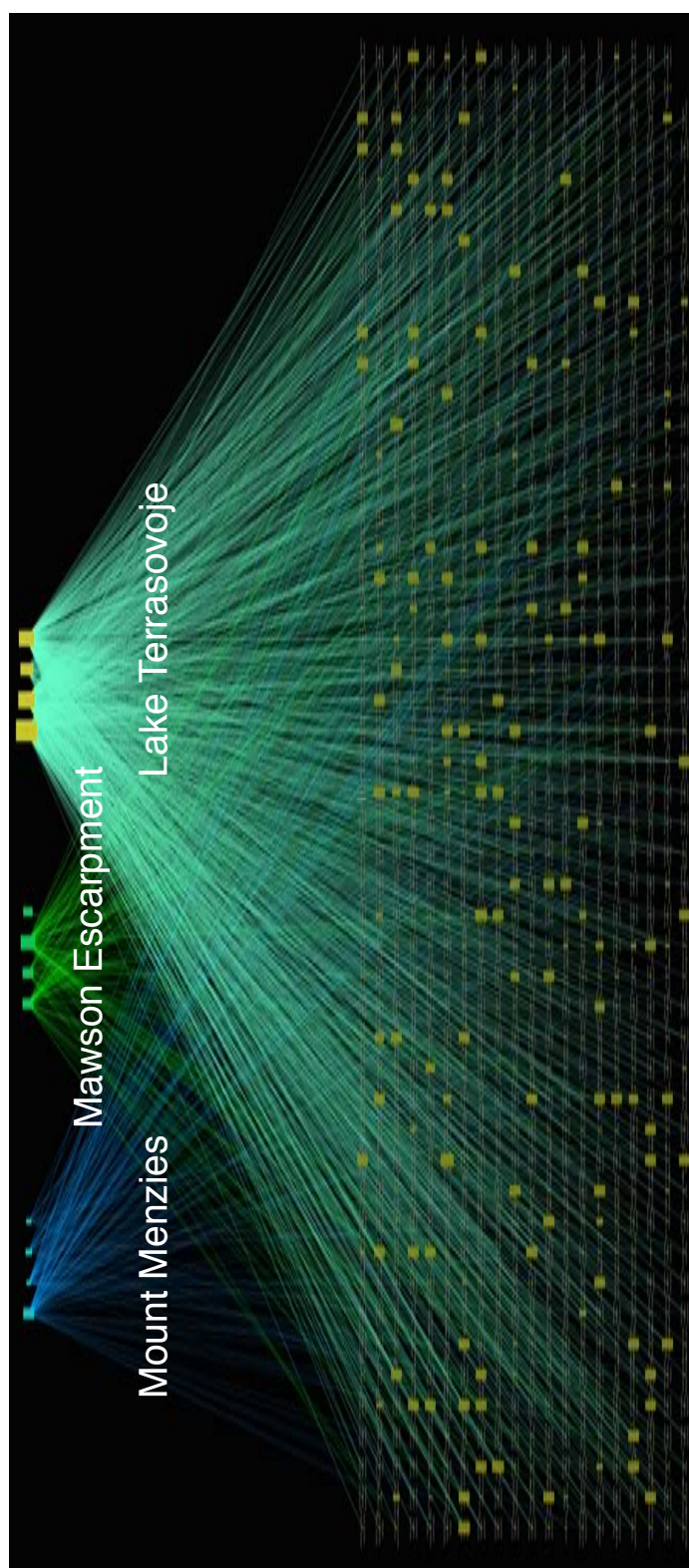


Figure 2.2.: Phylotype overlap between data generated for the four libraries generated from Mount Menzies, Mawson Escarpment and Lake Terrasovoje. Samples at the top: Mount Menzies, blue; Mawson Escarpment, green; Lake Terrasovoje, yellow. Phylotypes at the bottom: Square size reflects number of reads for each sample or phylotype. Figure created using un-rarefied data, coverage differences between the three locations are caused by low biological load at Mount Menzies and Mawson Escarpment when compared to Lake Terrasovoje.

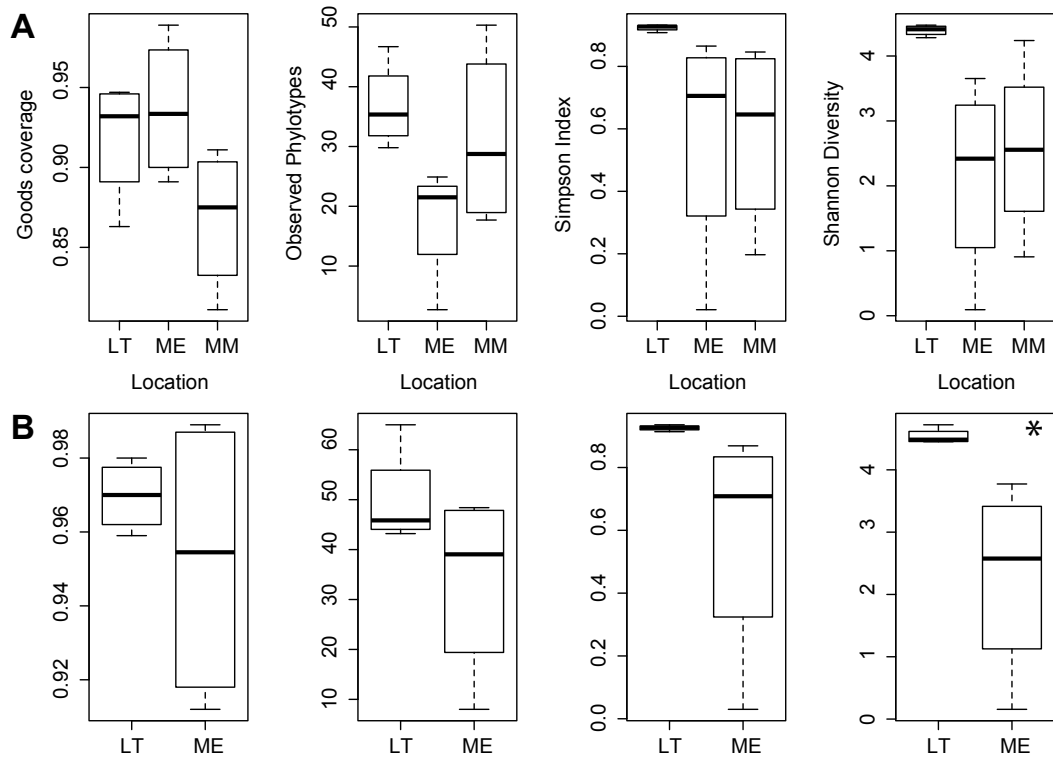


Figure 2.3.: Variance of α diversity metrics among regions at rarefaction depth of 160 (A) and 460 (B) sequences. Values in (A) are given for Mount Menzies (MM), Mawson Escarpment (ME) and Lake Terrasovoje (LT). Values in (B) are given for Mawson Escarpment (ME) and Lake Terrasovoje (LT). Significant differences only detected for Shannon diversity at depth 460 (*) between Mawson Escarpment and Lake Terrasovoje.

2.2.6. Species-level assignment of phylotypes

Species-level, location-specific phylotype composition was analysed using a second taxonomy assignment approach implemented in MEGAN v5.5.6 (Huson and Weber, 2013). While in QIIME taxonomic identities are assigned to clusters of sequences, in MEGAN, taxonomic identities are assigned to individual reads, and taxonomic assignments are subsequently offset against each other using a Lowest Common Ancestor (LCA) algorithm (Huson and Weber, 2013). Phylotype data were divided by region and sequences determined to be eukaryotic in QIIME were compared to the NCBI nucleotide collection (28/09/2014; BLAST v2.2.29, Altschul et al., 1990; using default parameters), and 50 results per query imported into MEGAN. Species-level taxonomy assignment was performed with strict LCA parameters to prevent spurious

matches (min. score 100, max. expected 0.001, top percent 10, min. support percent 0.1, min. support 1, LCA percent 99, min. complexity 0.44, use of minimal coverage heuristics). Taxonomic profiles were calculated with the projection method, but with further increased fidelity (min. support percent 1.0). Literature linked to the taxonomy assignments retrieved from NCBI was used to evaluate validity.

2.3. Results

2.3.1. Read processing

Despite equimolar pooling of libraries, un-rarefied, eukaryotic phylotype data was retrieved with unequal coverage across individual samples and sampling regions (Fig. 2.2). After quality filtering, 2 608 065 merged reads were screened for chimeras. From the remaining 2 607 945 reads, 8 126 (or 100 % of) phylotypes were obtained of which 6 779 (83 %) were assigned taxonomy and 5 449 (67 %) retained after removal of contamination using sub-setting functions in QIIME. After removal of non-eukaryote and unassigned phylotypes, a total of 2 403 (29.5 %) eukaryotic phylotypes were retained for further analysis. Eukaryote phylotype numbers were found to be lowest for Mount Menzies (73 °S; 3 330 m), intermediate at the Mawson Escarpment (73 °S; 807 m), and highest at Lake Terrasovoje (70 °S; 173 m, supplemental Tab. 2.4). The network analysis provided a graphical view of phylotype distribution per region; region separation (i.e. disconnected networks) was not observed (Fig. 2.2).

2.3.2. Eukaryotic α and β diversity comparison

Trajectories of rarefaction plots for all regions indicated comparable estimated sequence coverage, as observed by Good's indices from 0.8–0.9 (supplemental Fig. 2.6a). The number of observed species was lowest for Mawson Escarpment and highest for Lake Terrasovoje (supplemental Fig. 2.6b). Simpson and Shannon indices were comparable between Mount Menzies and Mawson Escarpment, and below the estimates for Lake Terrasovoje, (supplemental Figs 2.6c and 2.6d). Due to the low sequence count for Mount Menzies, we chose to exclude Mount Menzies from further analysis of diversity trends.

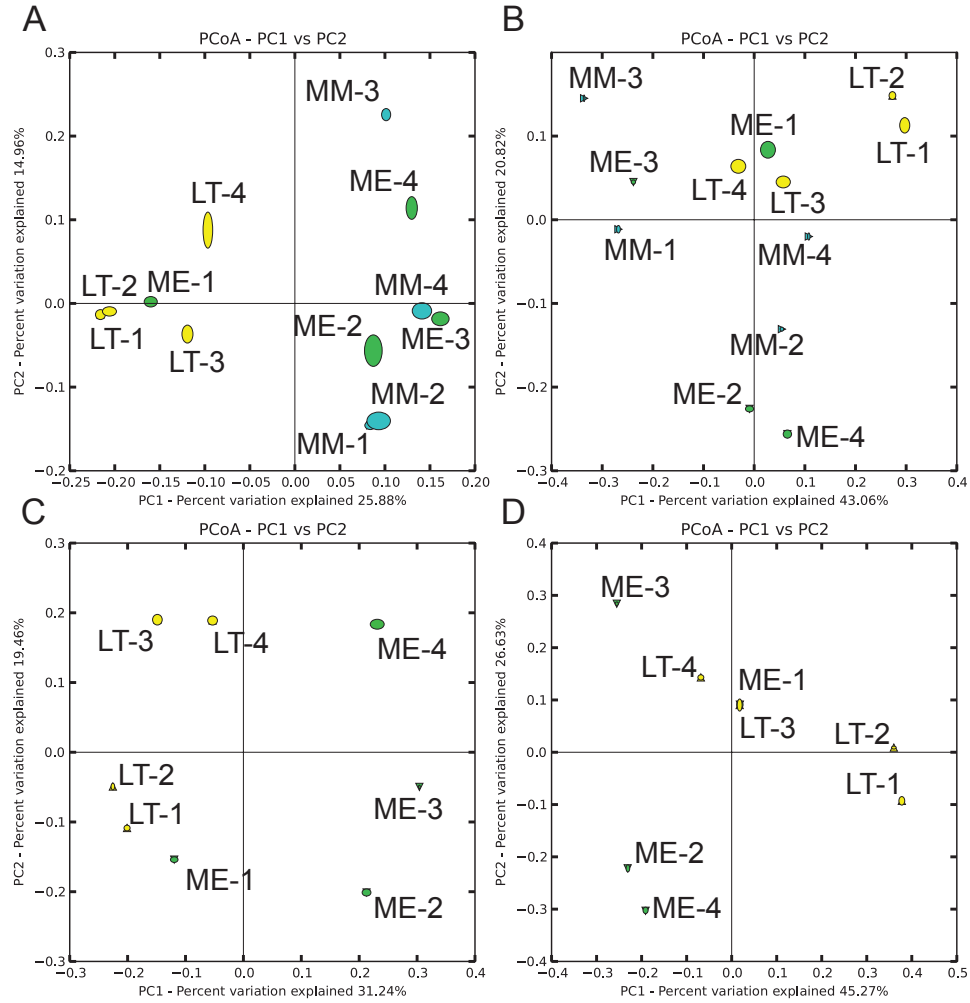


Figure 2.4.: Eukaryotic β diversity comparison for rarefaction depth of 160 (A, B) and 460 (C, D, without Mount Menzies) based on Principal Coordinate Analysis (PCoA) of unweighted (left; A, C) and weighted (right; B, D) UniFrac distances. Mount Menzies (MM): blue; Mawson Escarpment (ME): green; Lake Terrasovoje (LT): yellow. Ellipse size of samples represents error range based on Jackknife replication. Since sequence coverage was low for the Mount Menzies area, and significant differences in α diversity were only discovered when this sampling region was excluded from the analysis, we restricted our evaluation to differences between Mawson Escarpment and Lake Terrasovoje (panels C, D)

Estimates of α diversity measurements at the lower rarefaction depth (160 sequences) are provided in Fig. 2.3 a (for comparison) and reflect the trajectories of the rarefaction plots (supplemental Fig. 2.6) at the given rarefaction depth. Here, ANOVA estimated a non-significant influence of sampling location on Shannon diversity ($df = 2$,

$SS = 11.45$, $MSE = 5.723$, $F = 4.066$, $p = 0.055$), with Tukey's HSD's between Lake Terrasovoje and Mawson Escarpment ($q = 2.249$; 95 % CI $-z^* = -0.093$, $z^* = 4.591$; $p = 0.059$; not shown in Fig. 2.3 a, see supplemental Tab. 2.6). At the higher rarefaction depth (460; Fig. 2.3 b) median values for all four α diversity metrics were higher for Lake Terrasovoje compared to Mawson Escarpment. Significant differences were detected only for Shannon diversity (ANOVA: $df = 1$, $SS = 10.265$, $MSE = 8.297$, $F = 4.066$, $p = 0.028$; HSDs: $q = 2.266$; 95 % CI $-z^* = 0.341$, $z^* = 4.190$; $p = 0.028$; shown with asterisk in Fig. 2.3 b, also see supplemental Tab. 2.6).

Unweighted UniFrac distance comparisons of rarefied phylotype data grouped samples according to high or low numbers of phylotypes. This was obvious at the low rarefaction depth, for all samples (Fig. 2.4 a, LT-1 – LT-4, ME-1 versus MM-1 – MM-4, ME-2 – ME-4), and more pronounced for the regions Mawson Escarpment and Lake Terrasovoje (Fig. 2.4 c, LT-1 – LT-4, ME-1 versus ME-2 – ME-4). All samples with high read coverage were retrieved from altitudes below 600 meters and samples with low coverage were retrieved from higher altitudes (compare altitude values in supplemental Tab. 2.1). Weighted UniFrac distance comparisons, although not resulting in region-specific clustering, separated higher-altitude from lower altitude samples. This was indicated by the data from all three regions (Fig. 2.4 b, MM-2, ME-2 and ME-4 versus others) and more obvious when Mount Menzies was excluded from the analysis (Fig. 2.4 d, MM-2, ME-2 versus others; compare altitude values in supplemental Tab. 2.1).

2.3.3. Distribution of phylotypes across sites

No phylotype was present in more than six of the 12 samples. In six samples the most abundant phylotype was assigned to *Heterodermia boryi*, (mycobiont of desert ecosystems, see discussion) occurring at Lake Terrasovoje and two samples from Mawson Escarpment (Bonferroni corrected $p = 0.032$, 96.9 % sequence similarity, mean sequence counts 2340 and 79). In five samples the most abundant phylotype was assigned to Eimeriidae (bird parasites, see discussion) and other Apicomplexa and occurred at Lake Terrasovoje and the Mawson Escarpment (Bonferroni corrected $p = 0.031$, 96.7 % sequence similarity, mean sequence counts 900 and 67). Most abundant across four samples was a cercozoan (common soil eukaryote, see discussion) at Lake Terrasovoje (Bonferroni corrected $p = 0.038$, 97.7 % sequence similarity, mean sequence count 1520.75). The complete list of phylotypes is provided in Appendix A

and described in the supplemental information. Counts for phylotype groups (Fig. 2.5) detected across soil samples ranged from 1 to 29. The most widespread phylotypes were fungal, followed by non-algal protists. Algae showed a more limited distribution (i.e. were contained in three samples at most).

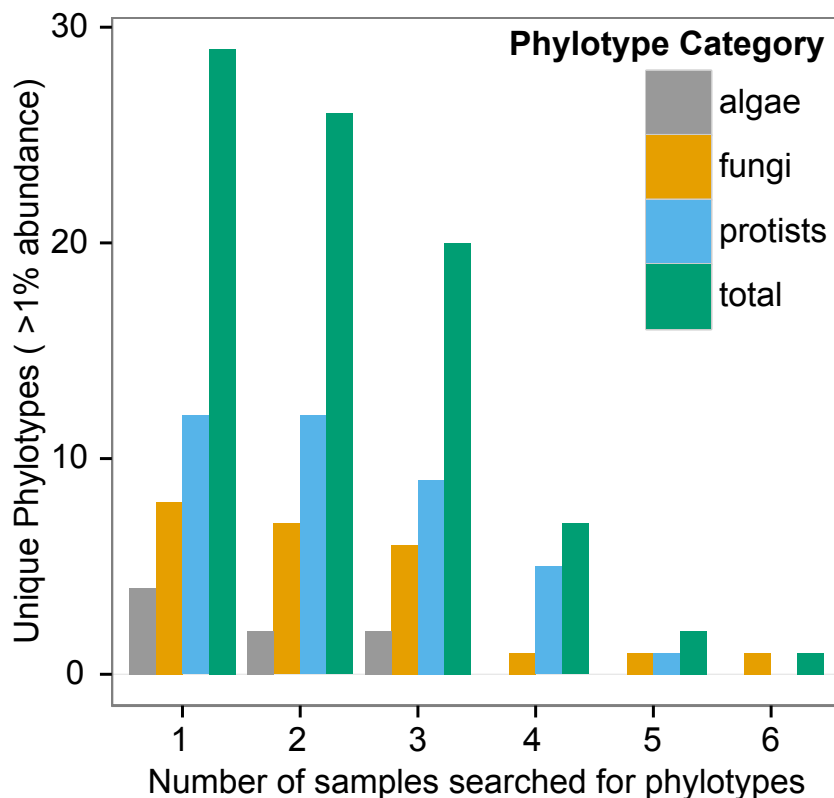


Figure 2.5.: Phylotypes detected across different numbers of samples, analogous to Lawley et al. (2004). Phylotype groups defined with phylotypes listed in supplemental information, restricted to phylotypes representing >1 % reads per sample.

2.3.4. Species-level assignment of phylotypes

Species-level assignments retrieved through MEGAN included Antarctic taxa or taxa likely to be present in Antarctica such as the lichen species *Lassalia pennsylvanica* (17 reads, Mawson Escarpment), the extremophile fungus *Cryomyces antarcticus* (128 reads, Lake Terrasovoje), the moss *Leptobryum pyriforme* (72 reads, Mount Menzies), the cercozoan *Cavernomonas stercoris* (21 reads, Lake Terrasovoje), as well as the green algae *Chlamydomonas reinhartii* (5 reads, Mawson Escarpment), *Characium perforatum* (7 reads, Mawson Escarpment), and *Koliella spiculiformis*

(13 reads, Lake Terrasovoje). The complete taxonomic profiles of all three sampling regions, along with literature evaluation are provided in supplemental Fig. 2.7 and supplemental information.

2.4. Discussion

Here, we (i) evaluated differences in eukaryotic diversity among three regions in the Prince Charles Mountains, (ii) analysed the identity and distribution of highly abundant phylotypes across samples, and (iii) examined the validity of species-level taxonomic assignments of Antarctic phylotypes using two different taxonomy assignment approaches. Our study provides a first view on micro-eukaryote diversity of the Prince Charles Mountains, and indicates a possible altitude and latitude related change of eukaryotic community composition between Mawson Escarpment and Lake Terrasovoje. Our work also serves as a guide regarding technical considerations for HTS metagenetic surveys.

2.4.1. Technical considerations

Unequal phylotype coverage among samples and regions (Fig. 2.2, supplemental Tab. 2.4) are not a result of extraction or PCR biases, but likely reflect the extremely heterogeneous distribution of Antarctic soil biota (Chown et al., 2015; Convey et al., 2014; Ettema and Wardle, 2002). Drastic diversity differences even between locations in close proximity are caused by different soil compositions and extreme fluctuations of abiotic conditions on small spatial scales (Bockheim, 1997; Convey, 2010; Magalhaes et al., 2012; Marchant and Head, 2007). Consequently, we increased the PCR cycle number to 35 in anticipation of low DNA yield from Antarctic soils (Dreesens et al., 2014), compared to other metagenetic assessments (e.g. 30 cycles in Bik et al., 2012). Subsequently, equimolar library pooling practically excluded PCR biases or extraction biases to be the cause of differences in phylotype coverage. Indeed, with the exception of MM-1, variation in the number of raw reads per sample was approximately one order of magnitude (supplemental Tab. 2.4), well within the range expected in HTS experiments. Hence, we believe the low number of eukaryotic reads from many of the Mount Menzies and Mawson Escarpment samples accurately reflects the limited amount of endogenous eukaryote DNA present in these samples. Additionally, environmental DNA can be highly degraded (Bellemain et al., 2013;

Taberlet et al., 2012), and soils with low biological activity (i.e. Mount Menzies, Mawson Escarpment) are presumed proportionally more abundant in degraded DNA than wetter and warmer soils (i.e. Lake Terrasovoje).

The validity of taxonomic placements of phylotypes is influenced by factors such as sequencing artefacts, contamination, marker choice, reference data and sequence coverage (Bohmann et al., 2014; Ficetola et al., 2014; Smith and Peay, 2014; Taberlet et al., 2012). We took appropriate measures to exclude sequencing artefacts. Our data filtering method (subtraction of OTUs present in PCR controls from Antarctic data) prevented us from analysing data that might constitute false positives, such as cryo- and halo-tolerant yeast species (Rao et al., 2012), which was advisable for the investigation of highly saline cold desert soils (Bockheim, 1997; Magalhaes et al., 2012) that we encountered at Mount Menzies and Mawson Escarpment. Although shortcomings of the 18S marker are known, its application allowed usage of the SILVA database and hence a comprehensive identification of eukaryotic organisms in the sampling region (Stenø ien, 2008; Tang et al., 2012; Zhan et al., 2014). Future environmental DNA studies of cold desert soils similar to Mount Menzies or the Mawson Escarpment will require higher sequencing effort to increase the strength of statistical conclusions on diversity trends. At the same time, the application of two taxonomy assignment approaches allowed detection of Antarctic eukaryotic phylotypes in the sampling area.

2.4.2. Differences in eukaryotic diversity among three locations

Although with weak statistical support, significant differences in Shannon diversity between Mawson Escarpment and Lake Terrasovoje are concordant with the general assumption that biodiversity increases with decreasing latitude and with less stringent environmental conditions on a global scale as well as in Antarctica (Gaston, 2000; Howard-Williams et al., 2006; Wu et al., 2011). Consequently, low biodiversity in extreme Antarctic environments (Wall and Virginia, 1999) such as Mawson Escarpment (and Mount Menzies) is responsible for the observed variation in sequence coverage. Since our observations are based on unrarefied data and only weakly supported in analysis of rarefied data, we limit conclusions to the observation that Lake Terrasovoje is richest and most diverse in soil eukaryotes in the investigated sampling area and that more southerly and higher altitude sites are likely to be substantially less diverse.

The UniFrac distance measure expresses the distance between communities based on the lineages they contain, the weighted UniFrac corrects these estimations for differences in sequence coverage (Lozupone and Knight, 2005; Lozupone et al., 2011). Consequently, groupings observed through unweighted distance analysis (Fig. 2.4 a, c) are indicative of high diversity and richness of phylotypes at Lake Terrasovoje and low-altitude samples from the Mawson Escarpment in comparison to other sampling regions. Weighted UniFrac distance showed each sample supported unique diversity based on local environmental factors (Fig. 2.4 b,d). More similar communities require more sequencing to reliably recover their relationships (Lozupone et al., 2011). Hence, differences between low-altitude and high-altitude samples from Lake Terrasovoje and Mawson Escarpment became more pronounced with increased sampling depth (Fig. 2.4 c, d). Our results are comparable to earlier metagenetic and traditional biodiversity studies in Antarctica that showed local biodiversity is highly variable depending on localised environmental factors (Lawley et al., 2004; Niederberger et al., 2015). A decrease of diversity with increasing altitude (i.e. from Lake Terrasovoje to the Mawson Escarpment) could either reflect general trends between biodiversity and altitude / latitude mentioned above, or age-related salt accumulation in dry polar desert soils encountered at Mawson Escarpment, when compared to Lake Terrasovoje, similar to observations in other parts of Antarctica (Bockheim, 1997; Magalhaes et al., 2012).

2.4.3. Distribution of highly abundant phylotypes

Through the application of HTS we show that protists and lichen (mycobiont) species constitute important components of Antarctic soil biodiversity. The widespread distribution of fungi including mycobiont species discovered here (Fig. 2.5) was also noted in Sanger based metagenetic studies of other regions of Antarctica (Fell et al., 2006; Lawley et al., 2004). Similarly, the general success of lichens was previously recognised for other regions of Antarctica (Kappen, 2000). In contradiction to earlier studies we find non-algal protists, rather than algae, to be the next most abundant and widely distributed phylotypes across samples (see Lawley et al., 2004), most likely due to the increased fidelity of our HTS based metagenetic approach in capturing soil diversity when compared to cloning approaches.

2.4.4. Validity of species-level taxonomic assignments of Antarctic phylotypes

Initial taxonomic placement of highly abundant phylotypes across the sampling range using QIIME allowed detection of organisms that would be expected in the sampling region. The most abundant phylotypes were assigned to *Heterodermia boryi*, Eimeriidae and Cercozoa. *Heterodermia boryi* is a mycobiont of desert ecosystems (Moberg and Nash, 1999), and hence the presence of related phylotypes is conceivable in the sampling region. Eimeriidae are apicomplexan bird parasites (Dolnik et al., 2009) and were only detected in locations where South Polar Skuas (*Catharacta maccormicki*) or Snow Petrels (*Pagodroma nivea*) were observed during field work (Lake Terrasovoje and Mawson Escarpment). Finally, cercozoans are known soil eukaryotes in Antarctica (Cavalier-Smith and Chao, 2003; Fell et al., 2006), and thus also likely to appear in the sampling region.

Using MEGAN, we implemented an alternative taxonomic assignment approach that combines taxonomic assignments from several individual reads using heuristic algorithms, while the QIIME environment assigns taxonomic identity to a phylotype combined from multiple sequences. Similar to QIIME, taxonomic assignments using MEGAN also resulted in the identification of species likely to appear in the sampling range. *Lassalia pennsylvani* is a lichen species reported from the Blue Mountains of Australia (Nash III, 1972). The presence of lichen is conceivable due to their dominance of Antarctic habitats (Kappen, 2000). *Cryomyces antarcticus* is a well-studied extremophile fungus known from the Ross Sea region (Selbmann et al., 2005). *Leptobryum pyriforme* is a moss species known from Dronning Maud Land (Kanda and Mochida, 1992). The algae *Chlamydomonas reinhartii*, *Characium perforatum* and *Koliella spiculiformis* are commonly found in Antarctic soils or may exhibit a ubiquitous distribution (Fell et al., 2006; Katana et al., 2001; Wilcox et al., 1993). Other, less likely, species-level assignments are likely to be caused by a current deficiency of reference data for the 18S gene region (Cowart et al., 2015). Consequently, increased availability of reference sequence data for Antarctic biota will allow better species-level taxonomy assignments in the future.

2.5. Summary and conclusions

We show that the application of HTS can provide a rapid survey of eukaryotic diversity in the PCMs. The relation between phylotype diversity and geographic location of the three sampling areas is likely indicative of increasing eukaryotic richness and diversity with decreasing latitude and altitude. In order to retrieve meaningful α diversity estimates for biologically depauperate Antarctic habitats using rarefied data, sequencing effort has to be further increased. At the same time, UniFrac distances were well suited to identify biological differences between both individual samples and regions. Through the application of HTS we were able to show that protists and lichen (mycobiont) species constitute important components of Antarctic soil biodiversity. Our HTS-based metagenetic approach demonstrates how taxonomic placement conducted using QIIME and MEGAN allows detection of highly abundant phylotypes that would be expected in the sampling region.

2.6. Acknowledgements

The Australian Antarctic Division provided funding under science project 2355 to M.S. The Australian Research Council supported this work through funds from linkage grant LP0991985 to A.C. and M.S. The University of Adelaide supported this project through the International Post-Graduate Research Scholarship to P.C. We thank the members of the field party, Tessa Williams, Fiona Shanhun, Adrian Corvino, Josh Scarrow and Nick Morgan. We are grateful for the support provided by the pilots of Helicopter Resources Pty. Ltd (TAS) for their invaluable support during the field campaign. We are indebted for the efforts of Perry Andersen, Michael Denton, and Bob Heath of Kenn Borek Air Ltd. in supporting our field campaign. We appreciate the help and support provided by the crew of Davis Station during the summer field season 2012. We appreciate the support of Alan McKay, Russell Burns, and the laboratory staff of the South Australian Research and Development Institute (SARDI). We thank the members of the Australian Centre for Ancient DNA for helpful comments on the manuscript and analyses. We appreciate the comments of two anonymous reviewers, who helped improve the manuscript considerably.

2.7. Supplemental data

Scripts and raw sequence data used for analysis are available via Digital Object Identifier 10.5281/zenodo.32030.¹

¹Data pre-release with closed access. Data will be available publicly after manuscript publication.

2.8. Supplemental information: Antarctic eukaryotic soil diversity of the Prince Charles Mountains revealed by high-throughput sequencing

2.8.1. Methods and Materials

2.8.1.1. Site description

Mount Menzies Mount Menzies is a peak at 73.5°S latitude and 61.83°E longitude with an elevation of 3355 m and located on the large massif between Mounts Mather and Bayliss, on the southern side of Fisher Glacier in the Prince Charles Mountains (U.S. Geological Survey – Geographical Names Information System, <http://geonames.usgs.gov>, retrieved 18.5.2014). The massif is composed of late Archean granite greenstone and part of the Ruker Terrane (Kamenev et al., 2009). Winds are strong and cold predominantly from the south-west (US national weather service, <http://earth.nullschool.net>, retrieved 23.5.14). Samples were collected from heights around 1830 m a.s.l from dry or snow-covered ground (see Tab.2.1). All samples were collected below the local ice advance of the last glacial maximum 14–9 kybp (White et al., 2011).

Mawson Escarpment The Mawson Escarpment is flat-topped, west-facing escarpment which extends in a north south direction for 113 km along the eastern side of the Lambert Glacier (U.S. Geological Survey – Geographical Names Information System, <http://geonames.usgs.gov>, retrieved 18.5.2014). The region is composed of late Proterozoic metamorphic rocks and part of the Lambert Province (Kamenev et al., 2009). Winds are moderate and predominantly easterly (US national weather service, <http://earth.nullschool.net>, retrieved 23.5.14). The Mawson Escarpment samples were collected from “Accidental Valley”, a dry valley uncovered from receded glaciers (White et al., 2011), from heights around 800 m a.s.l from slightly moist soils. All samples but one were collected below the local advance of the last glacial maximum 14–9 kybp (White et al., 2011).

Lake Terrasovoje Lake Terrasovoje is situated within the Amery Oasis, in an unglaciated area south of the Charybdis Glacier and east of the Loewe Massif (Wagner et al., 2004). Bedrock of the sampling area is composed of orthogneiss,

part of the tectonic province of the Beaver belt and covered with cenozoic moraine deposits (Kamenev et al., 2009). The area has the lowest altitude in the Prince Charles Mountains (Tingey, 1974). Predominant winds are southwesterly (US national weather service, <http://earth.nullschool.net>, retrieved 23.5.14). Samples were recovered from mostly moist soils (see Tab. 2.1). The last (local) glacial maximum receded in the area 18–12 kybp, considerably earlier than at the other sites (White et al., 2011).

Table 2.1.: Observation metadata for samples used for molecular analysis. Regions: “MM” - Mount Menzies, “ME” - Mawson Escarpment, “LT” - Lake Terrasovoje. Latitude and longitude provided in decimal degrees, datum WGS84. Elevation estimated using a handheld GPS receiver (Garmin eTrex Summit®, Lenexa, US-KS). Invertebrate counts based on 10 min observations.

Library	Region	Latitude	Longitude	Elevation (m)	Date	Time	Soil Temp. (°C)	Invertebrates	Slope (°)	Aspect
MM-1	MM	-73.4229	61.8712	1 513	26-Nov-11	11:00	4.1	0	5	NE
MM-2		-73.4003	62.0537	1 988	27-Nov-11	15:00	-15.3	0	3	SE
MM-3		-73.4258	61.9567	1 800	29-Nov-11	11:20	-4.3	0	10	NE
MM-4		73.4292	61.9403	2 020	29-Nov-11	12:45	-1.0	0	22	NE
ME-1	ME	-73.3263	68.3014	563	15-Dec-11	16:00	5.0	0	10	W
ME-2		-73.3047	68.4498	994	17-Dec-11	16:20	8.5	0	0	-
ME-3		-73.3164	68.4417	651	18-Dec-11	17:53	9.6	0	0	W
ME-4		-73.3091	68.4319	1 020	19-Dec-11	14:49	7.6	0	30	N
LT-1	LT	-70.5274	67.8257	103	13-Jan-12	12:30	9.1	116	2	NNE
LT-2		-70.5394	67.9112	223	13-Jan-12	15:05	8.1	0	0	-
LT-3		-70.5267	67.8723	156	13-Jan-12	15:30	13.7	0	8	SSE
LT-4		-70.5438	67.9061	212	13-Jan-12	18:31	11.9	30	0	-

Table 2.2.: Laboratory specific information for the generation of metagenomic libraries, incl. barcode assignment. Also provided are concentration values of DNA amplicons. Concentration of individual libraries on the TapeStation instrument estimated through integration under a peak of 250 bp.

Sample origin	Sample type	Extraction ID	Library ID	Barcode sequence	Amplicon conc. (ng / μ l)	Amplicon conc. (pmol / μ l)	Used in pooling (μ l)	In pool. (pmol)*
Mount Menzies	Soil	AC8310	MM-1, 131016AB1	AACAACCTGGCCA	1.82	1.08	4.12	4.5
Mount Menzies	Soil	AC8312	MM-2, 131016AB2	CGTGCACAATTG	0.53	0.31	14.15	4.4
Mount Menzies	Soil	AC8319	MM-3, 131016AB3	TAGCCGGAACCT	2.64	1.57	2.84	4.5
Mount Menzies	Soil	AC8320	MM-4, 131016AB4	CTCATCATGTTC	4.44	2.64	1.69	4.5
PCR Laboratory	PCR negative	-	131016AB5	TCAGGTTGCCCA	0.55	0.33	13.64	4.5
Mawson Escarpment	Soil	AC8343	ME-1, 131016AB6	CACCGAAATCTG	6.72	3.99	1.12	4.5
Mawson Escarpment	Soil	AC8347	ME-2, 131016AB7	CACGACTTGACA	7.64	4.54	0.98	4.4
Mawson Escarpment	Soil	AC8350	ME-3, 131016AB8	CGACACGGAGAA	0.9	0.53	8.33	4.4
Mawson Escarpment	Soil	AC8353	ME-4, 131016CD1	TACAGTTACGCG	4.98	2.96	1.51	4.5
PCR Laboratory	PCR negative	-	131016CD2	CGGCTAAGTTC	0.48	0.29	15.63	4.5
Lake Terrasovoje	Soil	AC8372	LT-1, 131016CD3	CATACAGCAACC	7.19	4.27	1.04	4.4
Lake Terrasovoje	Soil	AC8374	LT-2, 131016CD4	CCAGGGACTTCT	1.39	0.83	5.40	4.5
Lake Terrasovoje	Soil	AC8375	LT-3, 131016CD5	TCATTCCACTCA	1.1	0.65	6.82	4.4
Lake Terrasovoje	Soil	AC8377	LT-4, 131016CD6	TGACGTAGAACT	1.11	0.66	6.76	4.5
PCR Laboratory	PCR negative	-	131016CD7	CTTGGAGGCTTA	0.45	0.27	16.67	4.5
Other Antarctic Study	PCR positive	AV22	131016CD8	GAACCTATGACA	0.06	-	-	-

2.8.1.2. Sample selection

All extracts processed from Mount Menzies derived from samples collected in heights around 1 830 m a.s.l from dry or snow-covered ground and below the local ice advance

of the last glacial maximum 14–9kybp (White et al., 2011). Extracts processed from the Mawson Escarpment derived from samples collected in heights around 800 m a.s.l from slightly moist soils, all but one collected below the local advance of the last glacial maximum 14–9kybp (White et al., 2011). Extracts from Lake Terrasovoje derived from samples recovered from mostly moist soils. The local last glacial maximum receded in that area 18–12kybp, considerably earlier than at the other two sites, hence surfaces from Lake Terrasovoje constitute the oldest deposits in the sampling regime. Observations metadata of soil samples processed in this work are listed in Tab. 2.1. Sample details for molecular lab work are listed in Tab. 2.2.

2.8.1.3. Marker selection and primer structure

Like all other possible marker gene regions, the 18S gene region has drawbacks as well as benefits for the application in metagenomic studies. Firstly, the gene region is highly conserved and thus known to have impaired ability to resolve phylogenies in land plants (Stenøien, 2008), but was chosen for its ability to allow detection of a wider range of organisms than other gene regions (Medlin et al., 1988). Secondly, the 18S gene region is known to underestimate biodiversity in broad taxonomic surveys (Tang et al., 2012), but was chosen for offering extensive reference data, which is crucial for the detection of biodiversity. Also, the 18S gene region has frequently been applied in metagenomic surveys including Antarctica (Lawley et al., 2004; Venter et al., 2004).

We used the primers “Illumina_EukBr” as well as “Illumina_Euk_1391f” (the latter with different barcodes per sample, and sufficient redundancy to allow sequence error correction Golay 1949; Parfrey et al. 2014). These primers are suitable for paired-end 18S rRNA sequencing of eukaryotic communities on the Illumina MiSeq platform (Gilbert et al., 2010; Parfrey et al., 2014). “Illumina_Euk_1391f” contains a three domain priming sequence “1391f”. “Illumina_EukBr” carries a eukaryote-specific “EukBr” priming sequence (Medlin et al., 1988). Amplicon sequencing procedures for the MiSeq platform followed previous approaches (Caporaso et al., 2012; Parfrey et al., 2014).

Forward PCR primer sequence The forward primer contains the following sequences (Parfrey et al., 2014):

5' Illumina adapter - forward primer pad - Forward primer linker - Forward primer

(1391f).

5' AATGATACGGCGACCACCGAGATCTACAC - TATCGCCGTT - CG - GTACACACCGCCCGTC 3'

Reverse PCR primer sequences The reverse primer contains the following sequences (Parfrey et al., 2014):

Reverse complement of 3' Illumina adapter - Golay barcode (see Tab. 2.2) - Reverse primer pad - Reverse primer linker - Reverse primer (EukBr).

5' CAAGCAGAAGACGGCATACGAGAT - TCCCTTGTCTCC - AGTCAGTCAG - CA - TGATCCTTCTGCAGGTTACCTAC 3'

2.8.1.4. Amplification and library generation

Further comments on amplification PCRs were carried out in duplicate. Although three- to eightfold PCR replication recommended in metagenomic studies by some authors, sequence coverage constitutes the main factor that influences the credibility of ecological inferences (Ficetola et al., 2014; Gilbert et al., 2010; Smith and Peay, 2014). To allow cost-efficient processing, twofold PCR replication was chosen with regard to possible future larger-scale application. PCR of environmental DNA employing high cycle numbers is inadvisable, since it increases the amount of chimeric sequences (Ahn et al., 2012; Kanagawa, 2003). Even so, the comparatively high amount of PCR cycles was necessary to retrieve enough starting material from depauperate Antarctic soils. PCR products were pooled at equimolar ratios as listed in Tab. 2.2. We generated sequence data from PCR negative controls. Cross contamination during the extraction phase was controlled via concentration measurements of extraction blank reactions.

2.8.1.5. Technical sequence removal, clustering, taxonomy assignment

TRIMMOMATIC Quality filtering of metagenetic data is important to retrieve meaningful results (Bokulich et al., 2013). As a first step to achieve high quality data, forward and reverse reads were quality filtered and stripped from technical sequences (i.e. adapter and primer sequences) using TRIMMOMATIC 0.32 (Lohse et al., 2012) and a custom input *.fasta* file and parameters. The custom *.fasta* file for trimming contained the primer sequences as follows (newline character omitted):

>Prefix18S/1 CGGTACACACCGCCCGTC

```
>Prefix18S/2 CATGATCCTTCTGCAGGTTCACCTAC
```

```
>RevPrimerRC/1 GTAGGTGAACCTGCAGAAGGATCATG
```

```
>FwdPrimerRC/2 GACGGGCGGTGTGTACCG
```

TRIMMOMATIC was called in paired end mode according to the manual, and with the *clipping* and *filtering* parameters as follows:

```
ILLUMINACLIP:<filename>.fa:3:25:8:1:true MAXINFO:20:0.3 LEADING:4 MINLEN:
```

The ILLUMINACLIP parameter clips adapter and primer sequences contained in the specified *.fasta* file from both reverse and forward read with high accuracy. We allowed for three mismatches during sequence matching, using a threshold value of 25 to set the accuracy during palindrome clipping. Simple clip threshold was set to eight, the minimum adapter length for clipping was set to one bp. Both reads were retained for further processing (*true*) to retain quality values of both reads (as well as sequence differences in few cases). The MAXINFO parameter allows quality filtering of the clipped reads. With target length for the filtering algorithm set to 20, and a strictness of 0.3, we applied a moderate quality filtering at this stage. The LEADING parameter was used to clip low quality bases (below 4) from the beginning of the reads to facilitate read collapsing in further steps. MINLEN specifies that all reads below 30 bp in length were discarded, which facilitated collapsing of only high quality reads in the next steps. For further information please refer to the TRIMMOMATIC manual.

ADAPTERREMOVAL Corresponding read pairs that passed filtering through TRIMMOMATIC were collapsed using ADAPTERREMOVAL 1.1 (Lindgreen, 2012) ADAPTERREMOVAL was called in paired-end mode according to the manual and as follows:

```
--trimns --maxns 10 --trimqualities --minquality 10 --collapse --stats --minlength 50
```

```
--shift 20--pcr1 GTAGGTGAACCTGCAGAAGGATCA --pcr2 GACGGGCGGTGTGTAC
```

ADAPTERREMOVAL was chosen to collapse read pairs and to our knowledge currently² is the only program capable of collapsing paired end reads using quality scores and retaining quality scores in the collapsed reads. While primarily interested

²July 2014

in collapsing the paired reads into a single read and retain quality scores, we used the program to apply further quality filtering of the sequences. N's were removed from the read pairs using `--trimns`, the maximum number of N's per read was allowed to be 10 (`--maxns 10`). Quality trimming was applied to reads with a Phred score below 10 with `--trimqualities` and `--minquality 10`. Reads pairs were collapsed to single sequences (`--collapse`), the procedure was monitored (`--stats`), and the minimum length of reads to be kept was set to 50 (`--minlength 50`). To maximise the amount of collapsed reads the alignment between read pairs was allowed to be shifted a maximum of 20 bp (`--shift 20`). The adapter sequences were passed to the program only because it requires them as an input (`--pcr1 GTAGGTGAACCTGCAGAAGGATCA --pcr2 GACGGGCGGTGTGTAC`).

FastX toolkit The collapsed reads were quality filtered using the FastX toolkit 0.0.13 (<http://hannonlab.cshl.edu/>) to achieve an average quality of Q30 across at least 95% of each read:

```
Fastq_quality_filter -Q33 -q 30 -p 95 <filename>
```

Quality Assessment Read processing was monitored using FastQC v0.10.1 (<http://hannonlab.cshl.edu/>) and Geneious 7.1.2 (<http://www.geneious.com/>). Collapsed reads were converted from `.fastq` to `.fasta` using Galaxy (Giardine et al., 2005).

Chimera removal Artefact detection and removal is a frequent problem in metagenomic studies (Creer et al., 2010; Edgar, 2013; Ficetola et al., 2014). Chimera removal was done using QIIME environment 1.8.0 (Caporaso et al., 2010):

identify_chimeric_seqs.py identified putatively chimeric sequences among the unclustered sequences data by comparison to the SILVA database release 108 (http://www.arb-silva.de/no_cache/download/archive/qiime/ and Pruesse et al., 2007) with USEARCH (Edgar, 2010). The script was called with default parameters.

filter_fasta.py was used to remove chimeric sequences were removed from the un-clustered sequences data according to the QIIME manuals.

Sequence clustering and initial taxonomy assignment The validity of taxonomic placements of phylotypes is influenced by a number of factors, such as sequence artefacts, contamination and sequence coverage as well as marker choice, and most importantly, reference data (Bohmann et al., 2014; Ficetola et al., 2014; Taberlet et al., 2012). For these reasons was advisable to compare the QIIME based taxonomy assignment with another approach (MEGAN). Sequence clustering and initial taxonomy assignment was performed using the QIIME environment 1.8.0:

pick_otus.py was used for *de-novo* clustering of sequences at 97 % similarity employing UCLUST 1.2.22q (Edgar, 2010). The *de-novo* clustering approach was used to enable clustering unimpeded by potentially insufficient taxonomic information, the threshold was chosen to allow comparability to similar exploratory studies (Lawley et al., 2004). Reverse strand matching was enabled (**-z**), as well as the **--optimal** and **--exact** flags, to consider each sequence in the dataset as a possible cluster seed.

pick_rep_set.py with default parameters was used to extract cluster seed sequences from the complete data as representative phylotypes sequences for taxonomy assignment.

assign_taxonomy.py was then applied to assign representative phylotypes with taxonomy information from the SILVA database release 111 (ftp://thebeast.colorado.edu/pub/QIIME_nonstandard_referencedb/Silva_111.tgz) using UCLUST 1.2.22q. A similarity threshold of 90 % was chosen to assign phylotype sequences with taxonomy from the reference database.

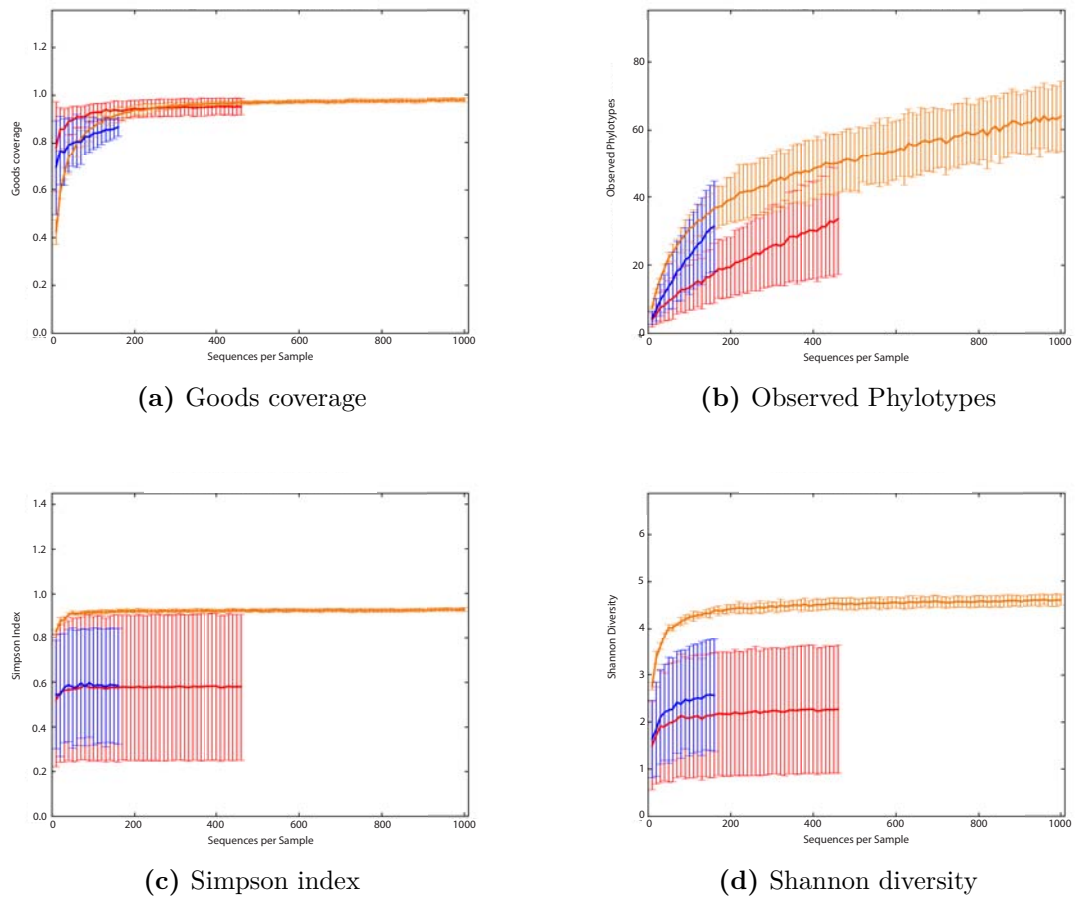


Figure 2.6.: Rarefaction curves of α diversity metrics evaluated for eukaryotic phylotypes of Mount Menzies, Mawson Escarpment and Lake Terrasovoje. (a) - Goods coverage, (b) - Observed Phylotypes, (c) Simpson index, (d) Shannon diversity Rarefaction conducted at depth between ten and 1000 sequences with increments of ten. Estimation of standard deviation from ten replicates per iteration. Further analyses were conducted at rarefaction depth of 160 and 460 sequences. Also see Table 2.5 for more details at employed rarefaction depths employed for further analysis.

2.8.1.6. Data analysis

Filtering Steps Data filtering was performed using QIIME 1.8.0:

`make_otu_table.py` recorded counts of phylotypes per library and taxonomy information in phylotype tables while with default parameters were applied.

`filter_samples_from_otu_table.py` and `filter_otus_from_otu_table.py` were used for removal of phylotypes without taxonomic information, subtraction of

phylotypes belonging to PCR negative controls and removal of non-eukaryotic phylotypes from the data. These steps removed a high number of phylotypes from the raw data, indicative of the successful removal of background contamination from the Antarctic eukaryote data (Tab. 2.4). Further information can be found in the QIIME documentation.

biom summarize -table provided summary counts and statistics of phylotype sequence abundance before and after filtering of phylotype tables. The **-qualitative** flag was employed for the retrieval of counts for unique phylotypes.

Eukaryotic α diversity comparison Eukaryotic α diversity comparison was performed using QIIME and R 3.1.1 (R Development Core Team, 2011):

make_otu_network.py was applied with default parameters and CYTOSCAPE 3.1.1, <http://www.cytoscape.org/>, to depict the un-rarified, quality-filtered, eukaryotic phylotypes.

alpha_rarefaction.py was used for eukaryotic α diversity comparison in conjunction and with parameter files to alter the default behaviour of the script. Phylotype tables were rarefied at depth from ten to 1000 sequences in step sizes of ten, with ten iterations per step to estimate rarefaction error. The script produced rarefaction plots for the chosen α diversity metrics for samples combined per sampling region (Mount Menzies, Mawson Escarpment and Lake Terrasovoje).

R v3.1.1 was used to compare eukaryotic α diversity between the three regions. Firstly, the averaged (from 10 iterations, see above) α diversity measures for each sample were obtained, namely values for both rarefaction levels and the four evaluated metrics (Goods coverage, Shannon diversity, observed phylotypes and Simpson index). These values were imported into R version 3.1.1 (2014-07-10), averaged per sampling location, and plotted using R's **boxplot()** function. Analysis of variance (ANOVA) was performed using the **aov()** function, followed by a Turkey's honest significant difference test (Tukey, 1949) using the **TukeyHSD()** function, the latter with and a confidence level of 95%.

Eukaryotic β diversity comparison

jackknifed_beta_diversity.py was used for eukaryotic β diversity comparison, using Unifrac distances between samples, rarefied data, and Principal Co-

ordinate Analysis (PCoA). After manually removing all gap-only sites of this alignment, alignment filtering was performed using the QIIME command `filter_alignment.py` with `-e 0.10` to remove the top 10% most entropic base positions, as recommended by the QIIME documentation. The script `jackknifed_beta_diversity.py` then calculated Unifrac distance metrics on rarified phylotype tables. Jackknife replicated was performed 160 (all three regions) or 460 times (excluding Mount Menzies). For further information refer to the QIIME documentation.

Distribution of phylotypes across sites

`group_significance.py` was used to detect unique phylotypes shared across increasingly smaller groups of samples belonging to the three sampling regions (Mount Menzies, Mawson Escarpment, Lake Terrasovoje), comparable to the approach of Lawley et al. (2004). Testing for significant differences in phylotype composition between sampling regions was performed using non-parametric analysis of variance (ANOVA-Kruskal and Wallis, 1952). One-thousand replicates were used for the non-parametric test.

2.8.2. Results and comments

2.8.2.1. Read processing

Technical sequence removal Tab. 2.3 provides an overview of the read counts during the read merging and filtering procedure. Quality filtering removed comparatively large amounts of reads from samples with low eukaryotic load, particularly from Mount Menzies, as obvious from the equimolar pooling of libraries (Tab. 2.2 and Tab. 2.4).

Table 2.3.: Results of read merging and quality filtering of raw data, as reported by FastQC. During sequencing a total of 74 libraries were sequenced, with 22 libraries related to this project (Tab. 2.2), and 52 samples from unrelated projects. The read count reported here only reports sequence counts related to this project.

Step	Raw reads		TRIMMOMATIC		Adapter removal	FastX filtering
	Mate 1	Mate 2	Mate 1	Mate 2	Merged	Merged
Read count	3,009,179	3,009,179	2,885,068	2,885,110	2,692,746	2,608,065
Mean quality per read	37	37	37	37	40	40
Sequence length	151	151	30–151	30–151	51–290	51–290

Chimera removal and successive phylotype filtering From a total of 2 608 065 merged reads, 611 were identified as chimeras based on reference comparison and 209 were identified as chimeras based on *de-novo* clustering. After removal of 120 chimeric sequences the union of sequences detected as non-chimeras by both methods (2 607 945) was used for clustering as per default setting in QIIME. Tab. 2.4 provides an overview of the phylotype counts during the filtering procedure. Subtraction of contamination contained in PCR negatives removed a high number of phylotypes from the raw data, indicative of the successful removal of background contamination from the Antarctic eukaryote data.

Table 2.4.: Phylotype counts during filtering steps, given are counts for sequences ('Reads') and counts of unique phylotypes ('Phylotypes') in each sample. First column contains amount of sequenced DNA for comparison, as also listed in Table 2.2. 'chimeras removed': Chimera-screened raw data; 'With taxonomy': data successfully assigned with taxonomy strings contained in SILVA database; 'Contaminants subtracted': counts after *subtraction* of data contained in PCR and extraction blank controls from the Antarctic samples; 'Eukaryotic phylotypes': Counts for data assigned to eukaryote domain. Phylotypes were clustered across all data and may overlap between regions (compare Fig. 2); hence summary counts are different to summary for complete data in main text.

Sample	Sequenced (pmol)	Chimeras removed		With taxonomy		Contaminants subtracted		Eukaryotic phylotypes	
		Reads	Phylotypes	Reads	Phylotypes	Reads	Phylotypes	Reads	Phylotypes
MM-1	4.5	3,201	76	3,088	75	164	52	163	51
MM-2	4.4	38,932	139	37,877	131	949	82	311	64
MM-3	4.5	46,472	294	31,063	190	8797	140	197	22
MM-4	4.5	64,501	420	61,908	390	6,794	337	2,375	143
Sum	-	153,106	929	133,936	786	16,704	611	3,046	280
Mean	-	38,276	232	33,484	196	4176	152	761	70
SD	-	25,726	155	24,200	137	4,272	128	1,077	51
ME-1	4.5	74,283	401	58,340	321	38,166	270	37,065	168
ME-2	4.4	33,535	161	31,322	112	1,353	84	628	62
ME-3	4.4	32,114	100	13,600	52	3,578	33	462	8
ME-4	4.5	45,111	318	40,664	281	8,668	226	1,175	92
Sum	-	185,043	980	143,926	766	51,765	613	39,330	330
Mean	-	46,260	245	35,981	191	12,941	153	9,832	82
SD	-	19,567	138	18,658	129	17,092	112	18,157	66
LT-1	4.4	47,981	697	41,543	619	25,741	547	19,159	144
LT-2	4.5	481,984	2,750	408,301	2,271	194,451	2,119	104,695	669
LT-3	4.4	509,729	2,567	454,680	2,235	143,266	2,091	60,432	767
LT-4	4.5	454,322	2,243	441,012	2,081	110,658	1,942	49,631	834
Sum	-	1,794,016	8257	1,345,536	7,206	474,116	6,699	233,917	2414
Mean	-	373,504	2,064	336,384	1,801	118,529	1674	58,479	603
SD	-	218,190	935	197,521	792	70,822	755	35,421	313

2.8.2.2. Data analysis

Eukaryotic α and β diversity comparison Results of rarefaction analysis, including error values are given in Tab. 2.5 for both rarefaction depths.

Table 2.5.: Eukaryotic α diversity metrics. Values given per sampling region, combined from four libraries per region, “MM” - Mount Menzies, “ME” - Mawson Escarpment, “LT” - Lake Terrasovoje. “n” = rarefaction depth, values are given for 160 and 460 sequences (“Goods” - Goods coverage, “Phylotypes” - observed phylotypes, “Simpson” - Simpson Index, “Shannon” - Shannon Diversity). Rarefaction results for Mount Menzies limited to a maximum of 160 sequences due to low sequence coverage for that sampling region. Error values (Standard Deviation) obtained from 10 iterations at the given depth. Observed species values rounded to integers.

Region	n	Goods		Phylotypes		Simpson		Shannon	
		\bar{x}	$\sigma_{\bar{x}}$	\bar{x}	$\sigma_{\bar{x}}$	\bar{x}	$\sigma_{\bar{x}}$	\bar{x}	$\sigma_{\bar{x}}$
MM	160	0.868	0.039	31	13.277	0.584	0.262	2.564	1.191
ME	160	0.937	0.039	17	8.745	0.574	0.331	2.146	1.321
	460	0.953	0.035	33	16.355	0.579	0.330	2.270	1.358
LT	160	0.918	0.034	37	6.245	0.924	0.009	4.395	0.073
	460	0.970	0.008	50	8.768	0.926	0.008	4.536	0.111

Results of ANOVA of region-specific α diversity variance means are given in Tab. 2.6. Differences of means are significant for Shannon diversity at both rarefaction levels, significance is stronger for the deeper rarefaction level. Variance differences of Shannon diversity are significant between Lake Terrasovoje and Mawson Escarpment for both rarefaction levels.

Distribution of phylotypes across sites All results of phylotype comparison across different sample groups in non-parametric ANOVA are shown in Appendix A. Phylotypes listed in that table are briefly commented here:

- The most widespread phylotype (six samples) was most prominent at Lake Terrasovoje, present at Mawson Escarpment and highly similar to *Heterodermia boryi*, a lichen mycobiont genus described from desert ecosystems (Moberg and Nash, 1999) (see main text and Appendix A).
- Undescribed representatives of the Apicomplexa (Eimeriidae) were assigned to a phylotype contained in 5 of the 12 samples predominantly at Lake Terrasovoje and also at Mawson Escarpment. Since closely related species are described as parasites in birds (Dolnik et al., 2009), and Mount Menzies was the only place where birds were not observed during the field campaign, the presence of Eimeriidae at Lake Terrasovoje and Mawson Escarpment could be linked to the presence of birds.

Table 2.6.: Results of pairwise comparison of region-specific α diversity metrics in ANOVA. Test were conducted for rarefaction levels of 160 sequences and 460 sequences, the former allowing inclusion of all libraries, the latter allowing testing of differences between Mawson Escarpment and Lake Terrasovoje at the deepest possible rarefaction depth. “Region” - Libraries were grouped into broader regions (Mount Menzies, Mawson Escarpment, Lake Terrasovoje) by obtaining the mean across all four respective samples, residuals are also given (“Residuals”). “Depth”: rarefaction depth, “Metric”: evaluated α diversity metric (“Goods” - Goods coverage, “Phylotypes” - observed phylotypes, “Simpson” - Simpson Index, “Shannon” - Shannon Diversity), “Df”: degrees of freedom, “Sum Sq”: sum of squares, “Mean Sq”: mean squares, “F value”: F ratio, “Pr (> F)”, significance probability value associated with the F Value. Significance codes: 0 – 0.001: “***”, 0.001 – 0.01: “**”, 0.01 – 0.05: “*”, 0.05 – 0.1: “.”, 0.1 – 1: “ ”. Significant values in **bold**.

Depth	Metric	Location	Df	Sum Sq	Mean Sq	F value	Pr (>F)	Significance
160	Goods	Location	2	0.01015	0.005073	2.731	0.118	
		Residuals	9	0.01672	0.001858			
	Phylotypes	Location	2	779.4	389.7	3.005	0.1	
		Residuals	9	1167.1	129.7			
	Simpson	Location	2	0.3171	0.15856	1.996	0.192	
		Residuals	9	0.7151	0.07946			
	Shannon	Location	2	11.45	5.723	4.066	0.0552	.
		Residuals	9	12.67	1.408			
	Goods	Location	1	0.000595	0.0005951	0.699	0.435	
		Residuals	6	0.005112	0.0008520			
460	Phylotypes	Location	1	534.6	534.6	2.329	0.178	
		Residuals	6	1377.4	229.6			
	Simpson	Location	1	0.2408	0.24082	3.317	0.118	
		Residuals	6	0.4357	0.07261			
	Shannon	Location	1	10.265	10.265	8.297	0.028	*
		Residuals	6	7.423	1.237			

- Other significant phylotype occurrences at Lake Terrasovoje were present across four soil samples and assigned to Thecofilosea and Glissomonadida, both are belonging to the cercozoans (Cavalier-Smith and Chao, 2003) and are known soil eukaryotes (Howe et al., 2011) (see Appendix A). Notably these cercozoan phylotypes are present in many more samples, although not in significantly different abundances between compared groups, most likely due to Bonferroni p values increased by larger samples sizes.
- Generally, between all compared groups of samples, Ascomycote and Basidiomycote phylotypes as well as protists and other unicellular eukaryotes determine site specific abundance differences, indicative of lichen as well as unicellular euk-

aryotes being the main components of biodiversity in the sampled regions. High diversity and widespread presence of lichen in the sampled regions corroborate views, which deem lichen to be greatly successful and diverse in Antarctica (Kappen, 2000). The presence of unicellular eukaryotes similar to the ones detected here (such Cercozoans and Stramenopiles) is known from early and recent clone-based metagenomic surveys conducted in Antarctica (Fell et al., 2006; Gokul et al., 2013; Lawley et al., 2004).

- A match to *Microcaeculus* (Acarina: Caeculidae) was detected at Lake Terrasovoje. The DNA sequence used for taxonomy assignment most likely stems from Australian specimens (Otto, 1993; Otto and Wilson, 2001) which only allow a coarse identification of this phylotype. Mites are well known from Antarctic soils (Convey et al., 2008) and different species were frequently observed during field work at Mawson Escarpment and Lake Terrasovoje.
- Furthermore, many eukaryotic phylotypes detected are flagged as uncultured, indicative of the presence of a not-well known diversity of soil eukaryotes in the sampling region.

Species-level assignment of phylotypes Comparison of taxonomic profiles at species level are shown in Fig. 2.7. Species level assignments within major detected groups there are briefly commented here:

Opisthokonts are by far the most diverse eukaryotic group detected in this work on a species level, in line with the overall high sequence coverage of this group. At the same time it is difficult to assess the validity of the species level assignments. In a few cases phylotypes are highly similar (i.e. within the threshold parameters of MEGAN, compare main text) to known Antarctic species, particularly when assignments have high query read counts. In many cases and when the query read count is low, assignments may be from organisms that are unlikely to appear in Antarctica. Among the latter are:

Alternaria alternata - An ascomycote known report from desert plant material in the USA (Parchert et al., 2012). Identified here from three reads derived from Mount Menzies

Sporidiobolales - Reported as yeast species from Thailand (Limtong et al., 2014). Antarctica is home to yeast species (Fell et al., 2006), but yeast species are also found anywhere else in the world.

Dreschlera biseptata - A plant pathogen of grass seeds, reported world wide (Leach and Tulloch, 1972). We did not observe grass species in the sampling area.

Toxicocladosporium posoqueriae - Fungus, plant pathogen, detected in Australia (Crous et al., 2012). Possible contamination since all lab work was carried out in Australia, or insufficient reference data.

Bagniseilla examinans - Identified on plant material also from deserts (Parchert et al., 2012). No higher plants present in sampling area.

Lophium mytelium - Fungus found on wood (Bisby and Dennis, 1952). Unlikely because no wood was observed in the sampling area.

Teratosphaeriaceae - Members of this family are likely to be contamination because members of this family are frequent pathogen of *Eukalyptus* (Pérez et al., 2012), common to Australia. Otherwise location bias in the reference data is possible.

Masseria sp. - Is a phylogenetically isolated genus of Ascomycetes, a plant parasite with high host specificity (Voglmayr and Jaklitsch, 2011). No higher plant were observed in the sampling area, hence occurrence of *Masseria* in Antarctica is unlikely.

The return of such unlikely hits could be caused either a lack of reference data, biased reference data, contamination, or chimeric sequences. Two phylotypes within the Opisthokonts were assigned to taxonomic groups whose presence is possible in Antarctica and passed the species assignment threshold in MEGAN. These are:

Lassalia pennsylvani - A lichen species reported from the blue Mountains (Nash III, 1972). Probable because lichen are common in Antarctica, despite the Australian species occurrence.

Cryomyces antarcticus - a well known and well studies Antarctic extremophile fungus from Antarctica, reported here in high abundance from Lake Terrasovoje.

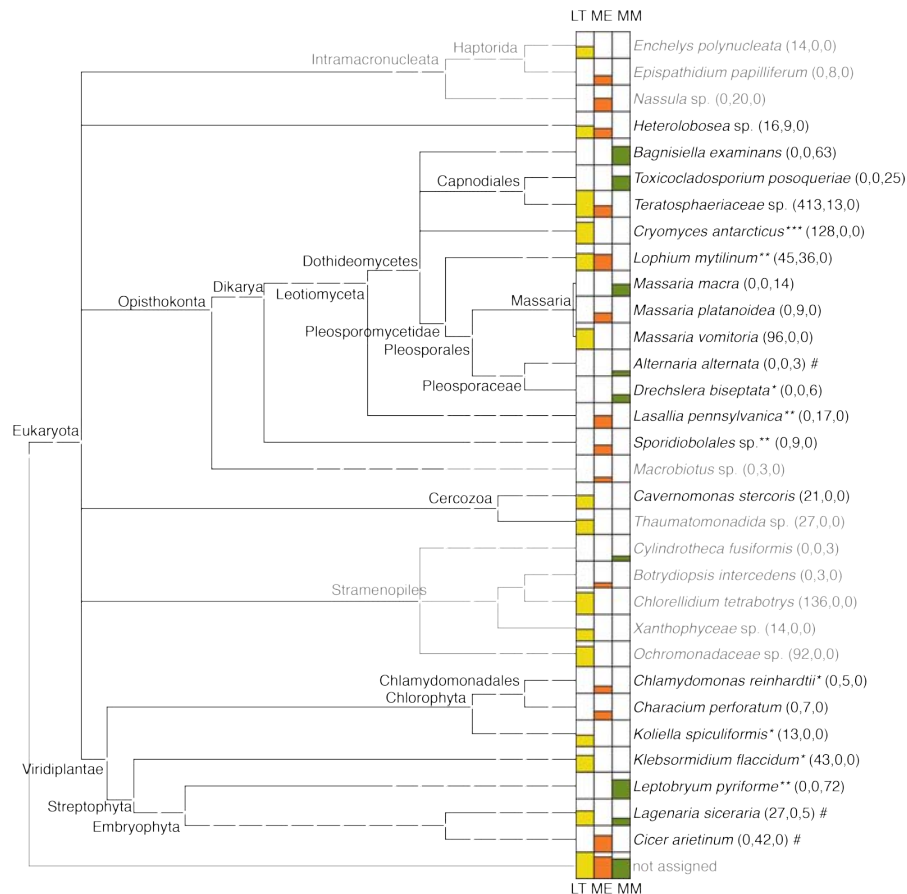


Figure 2.7.: Comparison of taxonomic profiles conducted in MEGAN for species-level assignment. Black - above profiling thresholds, grey - below profiling thresholds. Yellow - Lake Terrasovoje, orange - Mawson Escarpment, green - Mount Menzies. Bar-plots with logarithmic scaling, values in brackets indicate assigned reads per sample. Distribution of species checked via Global Biodiversity Information System (<http://www.gbif.org/>- 20.11.2014): Antarctic - “***”, Arctic and Subarctic “**”, temperate “*”. Likely misidentification or contamination - “#”. Interruptions of branches indicative of categories of the NCBI taxonomy, names are only given for selected categories.

Heterolobosea The match of phylotypes from Lake Terrasovoje and Mawson Escarpment to a single species of *Heterolobosea* could be indicative of these unicellular eukaryotes requiring water as well as certain other taxa to feed on - the only recent report mentioning *Heterolobosea* in Antarctic proximity reports several members of this groups feeding on bacteria and detritus at the coast of King George Island (Tikhonenkov, 2013).

Cercozoa are common eukaryotic soil predators that have been reported from the McMurdo dry valleys of Antarctica (Bass et al., 2009; Fell et al., 2006). The

taxonomy assignment identifies:

Cavernomonas stercoris - 4-9 μm in size, was first isolated from Antelope Island, Great Salt Lake (Bass et al., 2009). *Cavernomonas* is thus a relatively recently described cercozoan genus. Metagenomic surveys may in the future be well suited to detect microscopic eukaryotes across a wide range of possible new habitats, given that suitable reference sequences are available.

Streptophyta Phylotype identification appears to be highly dependent on read coverage per taxon, due to the LCA algorithm implemented in MEGAN (Huson and Weber, 2013):

Leptobryum pyriforme - Is a moss species known from Dronning Maud Land (Kanda and Mochida, 1992). While this assignment is a good example that the detection of Antarctic moss species is possible with the present methodology, it also becomes obvious that this is only possible with a sufficiently high read count. Streptophyte phylotypes with less coverage return spurious matches such as *Cicer arietinum* (chick pea) and *Lagenaria siceraria* (bottle gourd). Hence, while the 18S region appears to be well suited to detect algal phylotypes in soils (also due to the availability of reference sequences), the correct identification of mosses may be impeded by the slow evolutionary rate of the 18S marker in this group, that only allows its detection with sufficiently high read counts, due to the properties of the LCA algorithm (Huson and Weber, 2013; Stenøien, 2008). The detection of phylotypes similar to *Klebsormidium flaccidum* is interesting. Although this algal species has a recognised temperate distribution (see Fig. 2.7), it is also capable of cold acclimation (Nagao et al., 2008), and hence may in fact be present at Lake Terrasovoje.

Chlorophyta At Mawson Escapement and Lake Terrasovoje, species level assignments within Chlorophyta indicated the presence of phylotypes similar (i.e. within the threshold parameters of MEGAN, compare main text) to:

Chlamydomonas reinhardtii - Commonly found in soils (Marchant and Head, 2007).

Characium perforatum - Well studied soil algae (Wilcox et al., 1993).

Koliella spiculiformis - Well studied soil algae (Katana et al., 2001).

3. Matching phylotypes and morphotypes to invertebrate taxonomic assignments: implications for metagenetic surveys in terrestrial Antarctica

Paul Czechowski¹, Laurence J. Clarke^{1, 4, 5}, Alan Cooper¹, Mark I. Stevens^{2, 3}

¹ Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005 Australia; ² South Australian Museum, GPO Box 234, Adelaide SA 5000, Australia; ³ School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, SA 5000, Australia; ⁴ Australian Antarctic Division, Channel Highway, Kingston, TAS 7050, Australia; ⁵ Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, TAS 7001, Australia

Abstract Biodiversity information from Antarctic terrestrial habitats can assist studies estimating the effects of environmental change on Antarctic ecosystems, conservation management and investigating the historic effects of glacial constraints on the evolution and distribution patterns of Antarctic biology. Unfortunately, the distribution and diversity of the Antarctic biota to date is not well known, particularly in the case of micro-invertebrates. These invertebrates, such as springtails, mites, tardigrades, nematodes and rotifers, are morphologically conserved making identifications based on morphological approaches difficult. Not reliant on phenotype, identification using molecular methods may be better suited for the study of such taxa, but may lack resolution when sequence information is not used or may be prohibitively work intensive. Here, we compared the taxonomy assignment

performance of a high throughput sequencing metagenetic approach using one rDNA (18S) and one mtDNA (*cytochrome c oxidase subunit I* – COI) marker to reference data generated by morphological approaches. We were interested in how successfully each method (metagenomic or morphological taxonomic identification) retrieves taxonomic assignment on superphylum, phylum, class, order, family, genus, and species levels in an artificial DNA blend containing Australian invertebrates, and in seven extracts of *Antarctic soils* containing known compositions of microfaunal taxa. To avoid arbitrary application of metagenetic analysis parameters, we calibrated those parameters with metagenetic data from non-Antarctic soil extracts. We found that metagenetic approaches employing 18S and COI markers were well suited to detect the smallest and most cryptic Antarctic invertebrates, even when missed in morphological taxonomic assignments. On low taxonomic ranks 18S data outperformed COI data in accurately recognising Antarctic invertebrate phylotypes, likely due to lack of reference data for the COI marker with regard to Antarctic invertebrates.

Keywords Antarctica, invertebrates, environmental DNA, metagenetic, cytochrome c oxidase I, 18S rDNA

3.1. Introduction

Biodiversity information from Antarctic terrestrial habitats is important for estimating the effects of environmental change on Antarctic ecosystems (Freckman and Virginia, 1997; Nielsen et al., 2011), conservation management in light of increasing threats from non-indigenous invasive species (Chown et al., 2012), and investigations on the historic effect of glacial constraints on the evolution of Antarctic biology over millions of years (Convey and Stevens, 2007). Undertaking such biodiversity research in terrestrial Antarctica however, is challenging due to the logistics of accessing remote locations in a harsh environment (Convey, 2010). In recent years, biodiversity information for terrestrial Antarctic plant life has improved due to compilation of occurrence records from smaller-scale studies into easily accessible databases, and may in the future be easier to obtain through remote sensing technology (Fretwell et al., 2011; Peat et al., 2007). However, the distribution and diversity of Antarctic invertebrates remains largely unknown (McGaughan et al., 2011; Terauds et al., 2012) despite their important role in nutrient cycling and soil formation (Wall, 2012). Deficient biodiversity information for terrestrial Antarctic invertebrates is caused by

the persistence of slow and inefficient survey methods. Antarctic springtails, mites, tardigrades, nematodes and rotifers are morphologically conserved, but still frequently analysed with morphological approaches, requiring highly skilled taxonomists and ample time (Stevens and Hogg, 2003; Velasco-Castrillón et al., 2014). Not reliant on morphological identification, molecular methods are better suited for the study of such taxa, but may lack resolution when sequence information is not used (e.g. in analysis of Terminal Restriction Fragment Length Polymorphisms – TRFLPs) or may also be prohibitively work intensive, when large sample numbers are analysed (e.g. through Sanger-sequencing) (Makhalanyane et al., 2013; Nakai et al., 2012). *High Throughput Sequencing* (HTS) of amplicons generated from bulk extracts of environmental samples provides a more rapid generation of biodiversity information from terrestrial Antarctic habitats, which is deemed necessary for implementations of conservation approaches (Chown et al., 2015; Gutt et al., 2012). With such metagenetic methods, morphologically conserved species are rapidly distinguished in parallel using substrates such as soil, snow or water, while sampling procedures and laboratory workflows stay simple (reviewed in Bik et al., 2012; Bohmann et al., 2014). In Antarctica, HTS based metagenetic studies have investigated viruses (López-Bueno et al., 2009), bacteria (Bottos et al., 2014; Makhalanyane et al., 2013; Teixeira et al., 2010), eukaryotes (Dreesens et al., 2014; Niederberger et al., 2015) and could similarly be applied to invertebrates.

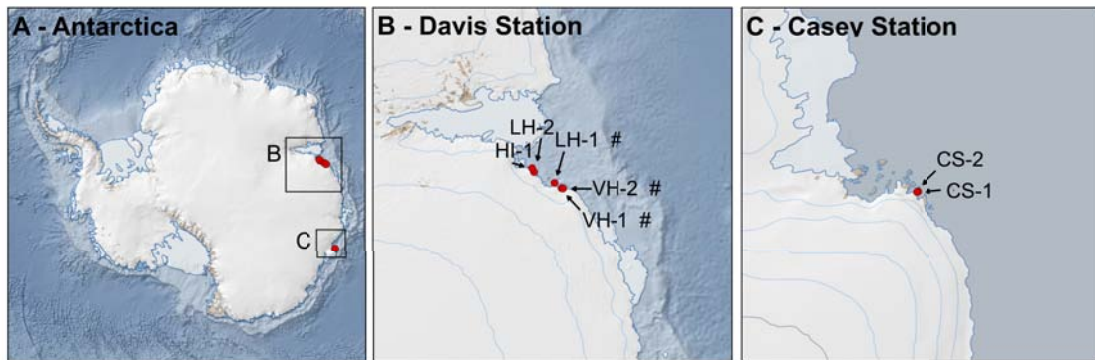


Figure 3.1.: Soil sampling locations used for morphological and metagenetic analysis of invertebrates. Amplification of 18S and COI metagenetic markers was conducted for whole-soil sample of all shown locations. From locations marked with a number sign data could only be retrieved using the 18S marker. Base layers compiled by the Norwegian Polar Institute and distributed in the Quantarctica package. Visit <http://www.quantarctica.org/>. Base layers courtesy of the SCAR Antarctic Digital Database, © 1993-2015 Scientific Committee on Antarctic Research; The National Snow and Ice Data Centre, University of Colorado, Boulder; NASA, Visible Earth Team, <http://visibleearth.nasa.gov/>; Australian Antarctic Division, © Commonwealth of Australia 2006.

It is currently unknown how well taxonomic assignments to Antarctic invertebrate phylotypes generated with metagenetic approaches would compare to taxonomic assignments of morphotypes. Metagenetic studies require suitable genetic markers to detect target organisms, and markers targeting 18S rDNA (18S) and cytochrome oxidase c I (COI) have been widely applied for phylogenetic studies of multiple invertebrate phyla which are also prevalent in Antarctica (Fell et al., 2006; Folmer et al., 1994; Lawley et al., 2004; Medlin et al., 1988). Both markers consequently offer a comparatively large amount of reference data to identify such invertebrates in mixed DNA extracts (Benson et al., 2011; Pruesse et al., 2007). A comparison of taxonomic assignments between 18S and COI phylotypes generated through metagenetic approaches and morphotypes would require evaluation across multiple taxonomic ranks and should consider available metagenetic reference data, and rank resolution of morphological identifications. Comparisons between morphological and HTS-based metagenomic approaches are also complicated by multiple assumptions regarding sequence clustering and taxonomy assignment. Often, analysis parameters for a given processing environment are more or less arbitrary, although crucial to establish reliable richness and diversity estimates (Koskinen et al., 2014; Smith and Peay, 2014).

Here, we compared the taxonomy assignment performance of a metagenetic approach using one 18S and one COI marker to reference data generated by morphological approaches. To avoid arbitrary application of metagenetic analysis parameters, we calibrated parameters with replicated metagenetic data from two soil samples (“Australian soils”). We were interested how successfully each method retrieves taxonomic assignment on superphylum, phylum, class, order, family, genus, and species level in an artificial DNA blend (containing Australian invertebrates - “*Australian blend*”), and in seven extracts of *Antarctic soils* (containing known compositions of microfaunal taxa - “*Antarctic soils*”).

3.2. Methods

3.2.1. Samples

Sampling locations of *Antarctic soils* are shown in Fig. 3.1, invertebrate isolation and taxonomic descriptions are detailed elsewhere (Velasco-Castrillón et al., 2014). Invertebrate morphotype composition of these soils is provided in Fig. 3.6. *Antarctic soils* were thawed and freeze-dried prior to DNA extraction. *Australian soils*, collected in Adelaide (July 2012, see supplemental Tab. 3.1), were introduced into the laboratory workflow at the freeze-drying stage; the *Australian blend* was introduced prior to amplification. The latter blend contained 15 taxa belonging to one order of Arachnida and 14 orders of Insects, at a total concentration of (3.1 ng/μl) (Clarke et al., 2014).

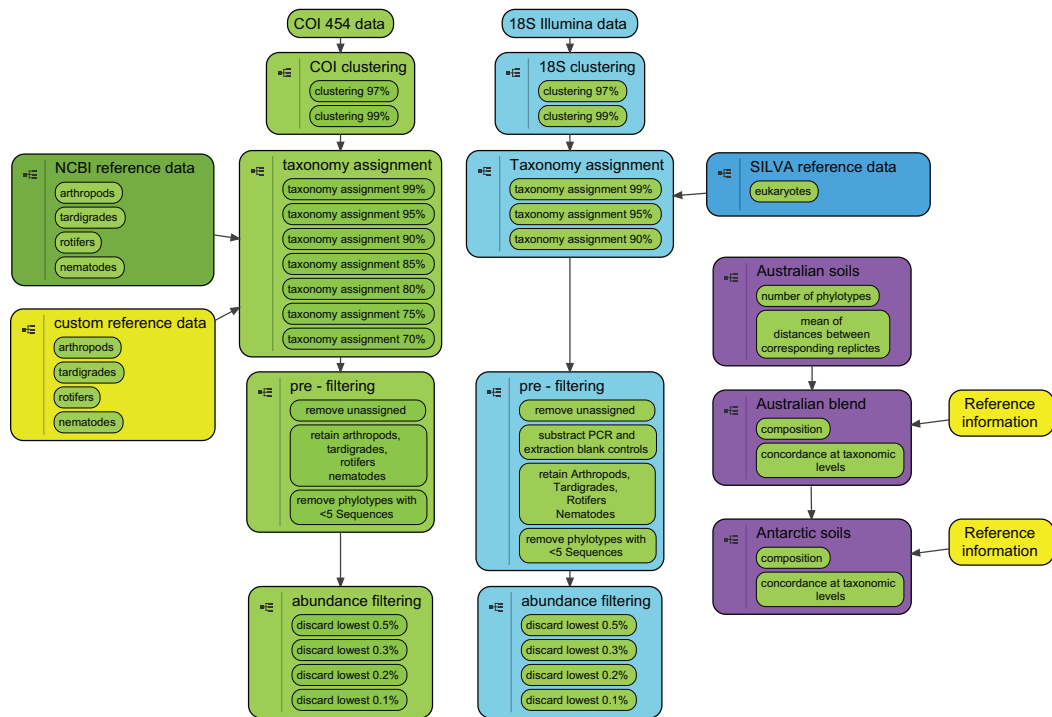


Figure 3.2.: Preparation of 18S (light green) and COI (light blue) phylotype data using the QIIME environment, and subsequent analysis (purple). During preparation, data of both metagenetic markers are independently clustered and assigned with taxonomy using multiple thresholds. Taxonomy assignment was aided by SILVA (Pruesse et al., 2007), NCBI (Benson et al., 2011) and unpublished reference data (18S, COI, respectively). During pre-filtering, phylotypes without taxonomic information or not defined by more than five sequences were discarded. Among the remaining phylotypes, only invertebrate phyla expected in *Australian blend* and *Antarctic soils* were retained for analysis. Subsequently, different percentages of low abundant phylotypes were discarded. During analysis (purple), comparisons of invertebrate phylotype compositions between two independent PCR replicates for each of two Australian whole soil extracts (*Australian soils*) were used to determine clustering, taxonomy assignment and abundance filtering parameters that yield similar compositions between corresponding replicates (without discarding all phylotypes). Those settings were then chosen to compare phylotype compositions of an *Australian blend* and seven *Antarctic soils* to their morphologic reference information.

3.2.2. DNA extractions

DNA extraction was performed at the South Australian Research and Development Institute (SARDI) using a method optimised for the retrieval of DNA from different

soil types and the retrieval of invertebrates in agricultural ecosystems for plant pathogen detection (Haling et al., 2011; Huang et al., 2013; Ophel-Keller et al., 2008; Pankhurst et al., 1996), that processes 400 g of starting material. Cross contamination during extraction was detected by measuring the concentration of blank extractions. DNA was stored at -20 °C (SARDI) and at -60 °C (University of Adelaide). Extraction of *Australian blend* is described elsewhere (Clarke et al., 2014).

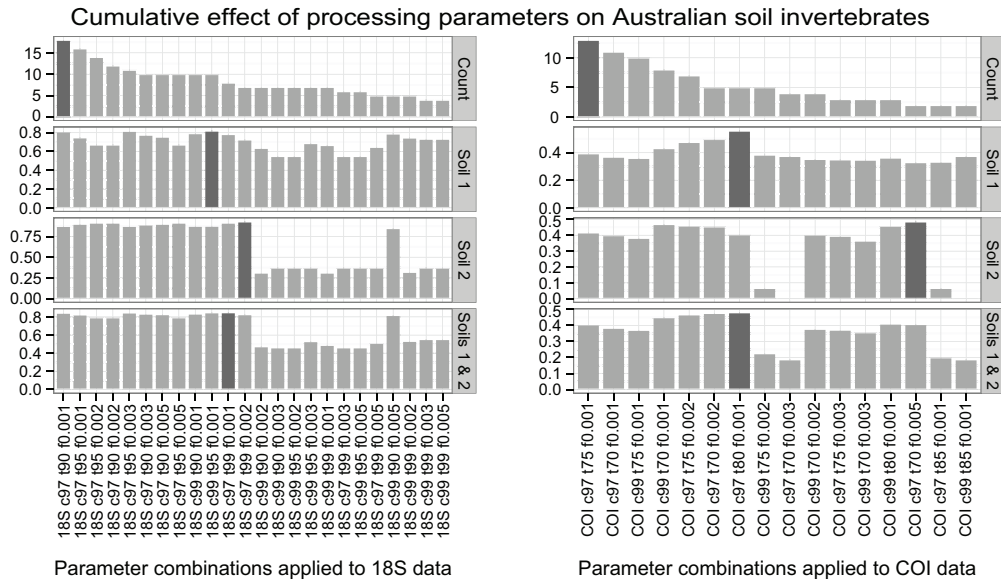


Figure 3.3.: Determination of suitable processing parameters (clustering, taxonomy assignment threshold, abundance filtering percentage) for invertebrate phylotypes recovered in *Australian soils* using 18S and COI metagenetic markers. 18S data on the left, COI on right, respectively. Count: Abundances of invertebrate phylotypes. Decreasing phylotype numbers increase compositional similarity between two independent PCR replicates of the same soil and thus were chosen to order processing parameters (in rows). Soil 1 / Soil 2: Jacquard indices described similarity between phylotype compositions of two corresponding PCR replicates for a given processing parameter and soil sample (Complete conformity = 1, complete nonconformity = 0). Soil 1 & 2: Mean Jacquard index was calculated from Soil 1 and Soil 2, which was used to choose a suitable processing parameter for each data set. Dark grey bars mark highest values in each row.

3.2.3. Primers

Primer sequences (including sequencing adapters and amplicon labels—*fusion primers*) for PCR and paired-end sequencing of 18S on the Illumina MiSeq platform were

sourced from the 18S rRNA amplification protocol 4.13 of the Earth Microbiome Project (Gilbert et al., 2010) and are also routinely employed by other groups specialised in metagenetic 18s rDNA analyses (Parfrey et al., 2014). Primers HCO2198 (Folmer et al., 1994) and mlCOIintF (Leray et al., 2013) were chosen for COI amplification and sequencing using the 454 GS FLX platform. Further details on fusion primer design for both gene regions are provided in sec. 3.8. Primer testing of 18S and COI fusion primers was performed; phyla Chelicerata, Nematoda and Rotifera could be recovered by both 18S and COI, phylum Tardigrada only by 18S.

3.2.4. Amplification and sequencing

Amplification and sequencing steps are detailed in sec. 3.8. Triplicate PCRs were prepared from all 8 extracts to alleviate mixed-template amplification biases (Bik et al., 2012; Bohmann et al., 2014; Gilbert et al., 2010; Kanagawa, 2003). Long extension times were chosen to counteract chimera formation (Lenz and Becker, 2008; Yu et al., 2012). Amplicons were visualised on agarose gels. Triplicates amplicons for each marker were then combined, purified and quantified. Amplicons above 0.25 ng/µl (supplemental Tab. 3.1) were then pooled by weight for each marker. Libraries were diluted to 9 pM for Illumina sequencing (18S) or concentrated to 3.18 ng/µl for emulsion PCR preceding 454 sequencing (COI). 18S libraries were paired-end sequenced in two separate runs on the Illumina MiSeq platform (Illumina, San Diego, US-CA) in 300 cycles and on two separate quarters of a 454 GS FLX PicoTiterPlate (COI). DNA extraction and PCR controls were included into amplification and sequencing for both markers, if the cleaned control reaction allowed pipetting (with a concentration above 0.25 ng/µl). Further details are provided in sec. 3.8.

3.2.5. Reference data for taxonomic assignments

For 18S taxonomy assignments, SILVA reference data (Pruesse et al., 2007) release 111 was used. Reference data for COI was compiled from earlier Antarctic studies (Velasco-Castrillón et al., 2014) as well as GenBank (Benson et al., 2011). Further details regarding creation and composition of reference data are provided in sec. 3.8.

3.2.6. Generation of phylotype observations using multiple parameter combinations

Phylotype data was generated in QIIME 1.8 (Caporaso et al., 2010), analyses were preformed in R 3.1.1 (R Development Core Team, 2011) using packages described elsewhere (Dray et al., 2007; McMurdie and Holmes, 2013; Wickham, 2007, 2009, 2011). With QIIME, we applied several clustering, taxonomy assignment and abundance filtering thresholds to metagenetic raw data of both markers and evaluated the effect of these different settings on phylotype data from *Australian soils* (Fig. 3.2), we then picked the most suitable setting to evaluate data from *Australian blend* and *Antarctic soils* (Fig. 3.3). Initially, deconvolution and chimera screening of 18S and COI data was performed. Subsequently, *de novo* clustering at 97 or 99% sequence similarity was performed with UCLUST (Edgar, 2010). Taxonomy assignment to phylotypes was performed with UCLUST and thresholds of 90%, 95% and 99% (18S), and 70%, 75%, 80%, 85%, 90%, 95% and 99% (COI; accommodating higher intraspecific pairwise distances between query and reference sequences). Resulting phylotype observations were filtered in a step-wise process (Fig. 3.2 and supplemental Tab. 3.3) to retrieve data free of information obtained from PCR and extraction blanks and containing only arthropods, nematodes, tardigrades and rotifers after removal of observations present at 0.1%, 0.2%, 0.3% or 0.5% total abundance. From 24 (18S) and 70 (COI) resulting QIIME phylotype tables, 24 and 16 contained data after processing and were imported into R using the Phyloseq package (McMurdie and Holmes, 2013). Morphological information for *Australian blend* and *Australian soils* was converted into a format accessible by Phyloseq and also imported into R. To ensure Antarctic phylotype origin in *Antarctic soils*, observations linked to *Australian soils* were removed from the *Antarctic soil* data. Taxonomy strings for morphological and metagenetic data were restricted or expanded (where possible), to yield superphylum, phylum, class, order, family, genus, and species rank-level information. Taxon names were corrected using NCBI taxonomy expressions (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>). All steps, QIIME-aided processing and R analysis are further detailed provided in sec. 3.8 and Appendix B.

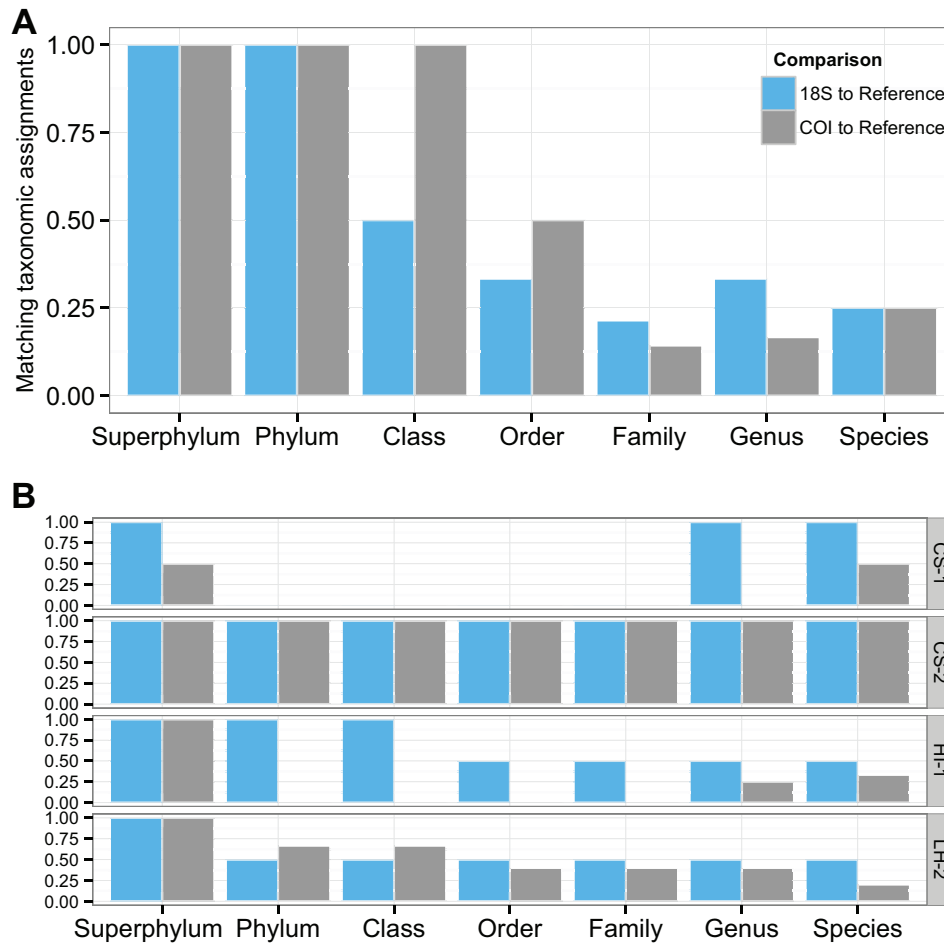


Figure 3.4.: Fraction of detected taxonomic entities in metagenetic data, when compared to morphologic reference information at various taxonomic levels. A: *Australian blend* B: *Antarctic soils*. Blue: 18S. Grey: COI. Inclusion of unassigned taxonomy in both reference and metagenetic data increased similarity on lower taxonomic levels even when higher taxonomic levels are not concordant (e.g.: sample CS-1).

3.2.7. Selection of processing parameters for 18S and COI phylotypes

We selected QIIME processing parameters (clustering threshold, taxonomy assignment, low-abundance filtering percentage) based on the highest mean value of Jacquard indices between individually replicated PCRs of each of two *Australian soils* (Fig.3.3). The chosen analysis parameters retrieve the most compositional similarity between two PCR replicates of two *Australian soils* and minimise inclusion

of low abundant phylotypes without discarding phylotypes reflective of the ‘true’ compositional diversity (see discussion, on page 72).

3.2.8. Concordance between taxonomic assignments of morphotypes and 18S / COI phylotypes

To compare the taxonomic resolution of morphotype assignments in *Australian blend* and *Antarctic soils*, in order to evaluate the quality of taxonomic assignments to morphotypes, we plotted their taxonomic composition among the lowest completely defined taxonomic ranks (family level for *Australian blend* and order level for *Antarctic soils*). Concordance between taxonomic assignments of morphotypes and 18S / COI phylotypes was determined by subsequently comparing rank-level information obtained for morphotypes and phylotypes across all (seven) taxonomic ranks. At each rank level, all occurring rank names contained in the morphologic data were recorded, and their presence was evaluated among phylotypes obtained for each gene. Complete concordance between morphotypes and phylotypes was assigned value “1”, complete dis-concordance was expressed through “0”. In order to not deflate taxonomic concordance when both morphological and metagenetic data did not yield taxonomic information for specific ranks, unavailable taxonomic information was not excluded, but coded as unavailable (“NA”). To compare taxonomic assignment correlations between 18S and COI phylotypes in relation to morphotypes in dependence of the used extract concentrations (i.e. high, *Australian blend*, versus low, *Antarctic soils*, see supplemental Tab. 3.1) and available morphotype information (defined to lower ranks, *Australian blend*, versus defined to higher ranks, *Antarctic soils*; see supplemental Fig. 3.7 and Fig. 3.8, respectively) we calculated inter-class correlation coefficients (ICC) (Koch, 1982). These ICCs related the 18S and COI concordance values, when each marker was compared to the morphologic reference data, and reach a value of 1 if both markers have the same ability to detect morphotypes. R source code for analysis is provided in Appendix B.

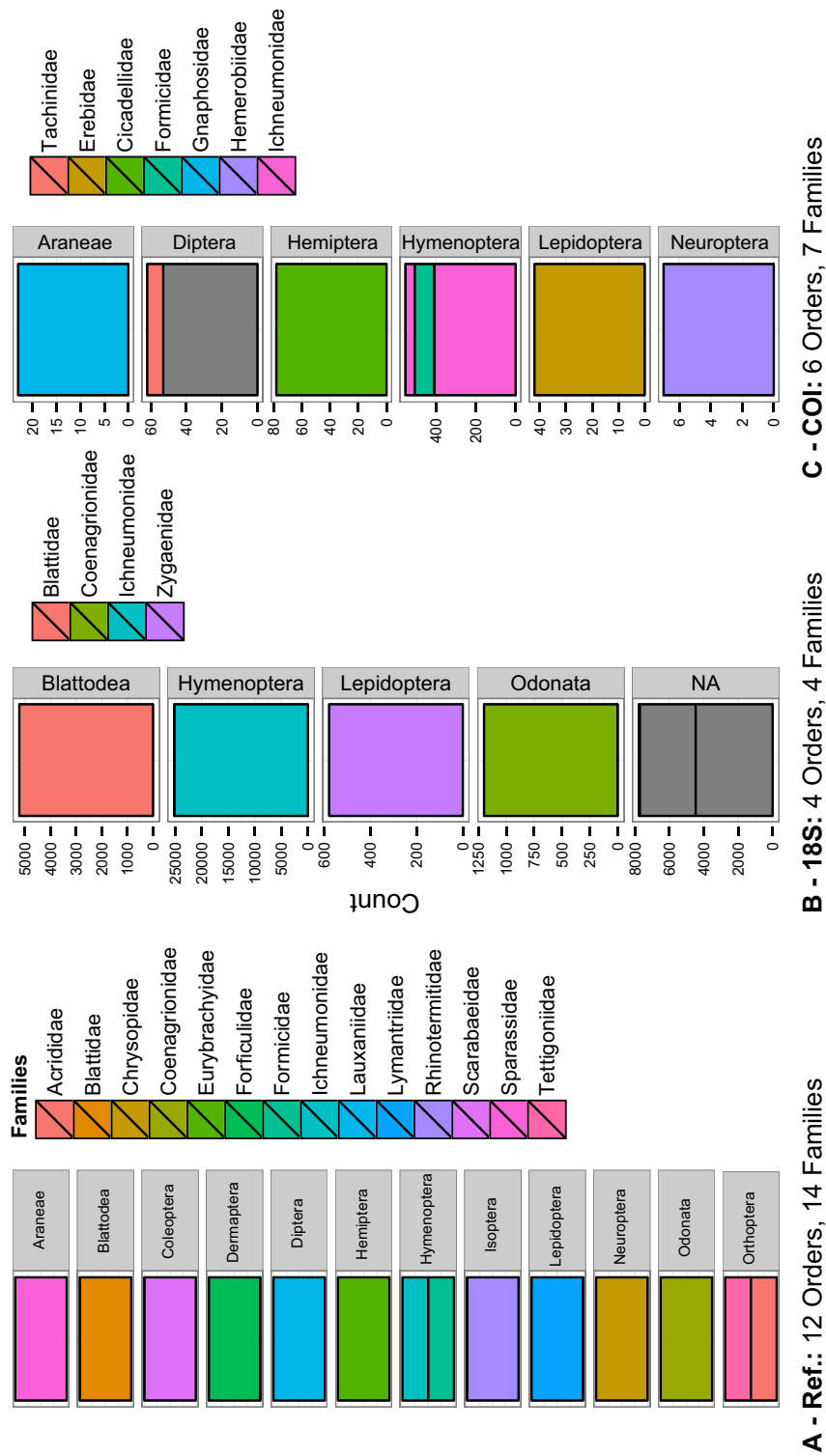


Figure 3.5.: Taxonomic assignments to invertebrate contained in *Artificial blend*. Composition is shown on order and family level.

3.3. Results

3.3.1. Selection of analysis parameters for 18S and COI metagenetic data

Maximum mean compositional similarity between two PCR replicates of each of both *Australian soil* samples (with 0.8 for 18S, and 0.45 for COI) was achieved using a clustering threshold of 97%, a taxonomy assignment threshold of 99% and using low abundance filtering of 0.01% for 18S. Values for COI were determined with 97%, 80% and 0.01%, respectively (Fig. 3.3).

3.3.2. Concordance between taxonomic assignments of morphotypes and 18S / COI phylotypes

For *Australian blend*, the ICC coefficient of 18S and COI phylotype taxonomy assignments in relation to morphotype taxonomy was 0.84, which served as a comparison value to *Antarctic soils* (supplemental Tab. 3.4). Phylotype taxonomy assignments of *Australian blend* of 18S and COI were accurate on the superphylum and phylum level (Ecdysozoa and Arthropoda, respectively, Fig. 3.4a). On class and order levels COI performed better than 18S (Fig. 3.4a) - insects were more accurately retrieved by COI, furthermore six of 12 expected orders were retrieved (Araneae, Diptera, Hemiptera, Hymenoptera, Lepidoptera, Neuroptera, Fig. 3.5c). 18S yielded only four expected orders (Blattodea, Hymenoptera, Lepidoptera, Odonata, Fig. 3.5b). On the family level 18S yielded higher matches (Fig. 3.4a), three of 12 expected families were accurately assigned (Blattidae, Coenagrionidae, Ichneumonidae) and one family (Zygaenidae) constituted a miss-assignment (Fig. 3.5b). In comparison, COI yielded only two correct family assignments (Formicidae, Ichneumonidae), while five families were miss-assigned (Tachinidae, Erebidae, Cicadellidae, Gnaphosidae, Hemerobiidae, Ichneumonidae, Fig. 3.5c). The concordance indices hereafter rose for both markers on the genus and species level (Fig. 3.4a), indicative of missing taxonomic information in both the morphologic and metagenetic data (Fig. 3.5a and supplemental Fig. 3.7). Morphotype taxonomy assessments performed comparatively well for relatively large invertebrates and were possible in *Australian blend* to species rank (Fig. 3.5a and supplemental Fig. 3.7).

In the *Antarctic soils* samples ICCs for samples CS-1, CS-2 and HI-1, were 0.42, 1.0

and -0.73 respectively (supplemental Tab. 3.4), and thus lower than for the *Australian blend* (CS-1, HI-1) and / or influenced only by unavailable taxon information (HI-1) (Fig. 3.6, supplemental Fig. 3.8). Sample LH-2 yielded a comparatively high ICC value (0.76), also due to the detection of Plectidae (Araeolaimida, Nematoda) in all three data sets (Fig. 3.6). In this sample, COI performed better on phylum and class level, 18S performed better on order to species ranks (Fig. 3.4b). Morphotypes across all *Antarctic soils* comprised of taxa in six orders (Adinetida, Araeolaimida, Rhabditida, Dorylaimida, Parachela and Phylodinidae) and seven families (Macrobiodidae, Adinetidae, Hypsibiidae, Philodinidae, Plectidae, Qudsiannematidae and Rhabditidae, Fig. 3.6a). 18S data yielded only one of those orders and families (Araeolaimida: Plectidae), in 3 of 5 expected samples LH-1, LH-2, VH-1 (Fig. 3.6b). Furthermore, 18S phylotypes comprised of orders not detected in morphologic approaches (Monhysterida in sample HI-1 and Oribatida in sample LH-2, Fig. 3.6b), each comprised of one family (Monhysteridae and Phenopelopidae, respectively). COI metagenetic data yielded two orders contained in morphologic reference data (Adinetida and Araeolaimida, Fig. 3.6c). COI family level assignment to Araeolaimida was concordant with morphologic data (Plectidae, Fig. 3.6a,c). This family was detected in sample LH-2 with both approaches, in CS-1 only with COI (Fig. 3.6a,c). In order Adinetida, morphologic assessment yielded the family Adinetidae in sample LH-1 and LH-2 (Fig. 3.6a), for which matching information was unavailable in COI (Fig. 3.6c); instead family Adinetidae was detected in sample CS-1 (Fig. 3.6c). Since we excluded non-Antarctic phylotypes in our initial processing and also conducted chimera filtering and low-abundance filtering, orders Coleoptera, Diptera, Lepidoptera and families therein (Fig. 3.6c) are highly likely to constitute miss-assignments due to missing reference data. When compared to *Australian blend*, taxonomy assignment comparisons between phylotype and morphotypes were more impeded by missing information; determination of morphotypes became difficult below order level for these morphologically highly conserved invertebrates (Fig. 3.4b, Fig. 3.6 and supplemental Fig. 3.8).

3.4. Discussion

We examined the usefulness of two gene marker regions with comparatively comprehensive reference information for application to Antarctic invertebrates in a metagenetic approach. In doing so, we add to the range of metagenetic studies that

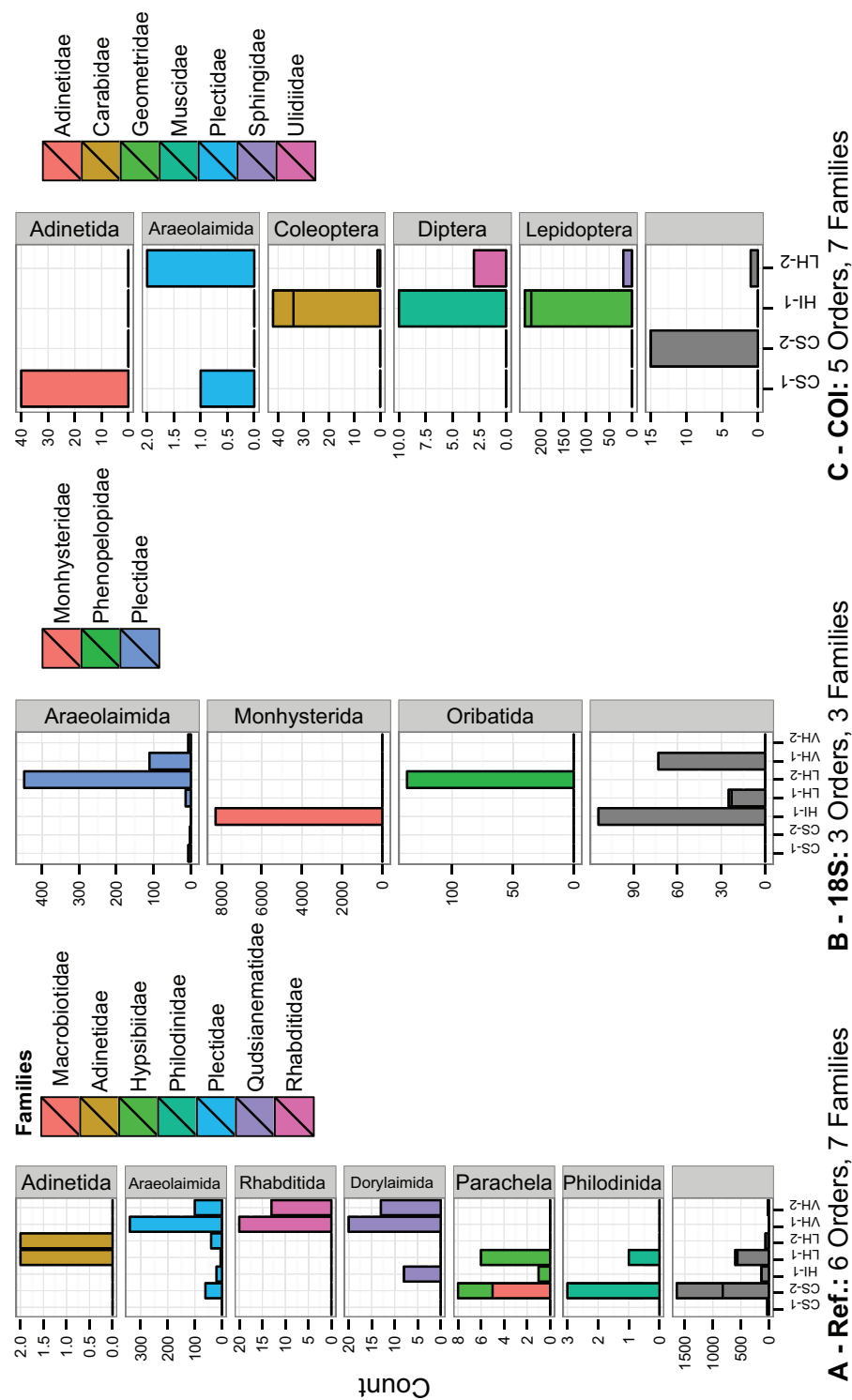


Figure 3.6.: Taxonomic assignments to invertebrate contained in *Antarctic soils*. Composition of morphotypes and phylotypes is shown on order and family level.

compare metagenetic markers (e.g. Zhan et al., 2014; Zhang and Hewitt, 1997), focus on invertebrates (Cowart et al., 2015; Tréguier et al., 2014; Wu et al., 2009) and provide replicated ‘ground truthing’ with morphological data (Cowart et al., 2015; Tréguier et al., 2014). We also expand the range of studies investigating the effect of analysis parameters on metagenetic data sets (Bik et al., 2012; Bokulich et al., 2013; Koskinen et al., 2014; Zhan et al., 2014) including the quality of reference data (Berney et al., 2004; Kwong et al., 2012). We show that both metagenetic data generation and morphologic taxonomy assignments are more complicated for small, cryptic Antarctic terrestrial invertebrates than for larger, less cryptic invertebrates of higher diverse and taxon-rich ecosystems, where also comparatively more reference data exists. Our approach allows detection of Antarctic phylotypes that are not contained in metagenetic reference data, as discussed below. Metagenetic markers employed for the detection of Antarctic invertebrates from bulk soil extracts need to be chosen depending on the desired rank-resolution of taxonomic assignments, and with regard to available reference data. On lower taxonomic ranks, the taxonomic resolution of 18S currently outperforms that of COI in metagenetic approaches applied to Antarctic invertebrates. Comparable results regarding the low rank- resolution of 18S and COI might not apply to low-diversity terrestrial ecosystems other than Antarctica, if more comprehensive reference data is available for the COI marker.

3.4.1. Selection of analysis parameters for 18S and COI metagenetic data

Previous studies demonstrated that amplicons of replicate bulk soil DNA extracts yield similar taxonomic compositions when the same markers are used, even when sequenced on different HTS platforms (Smith and Peay, 2014). At the same time, low abundant phylotypes in metagenetic data are more likely to be of chimeric origin or may constitute PCR or sequencing errors and their removal may improve ecological inferences (Bokulich et al., 2013). Sequence data processing thresholds adjustment (for clustering, taxonomy assignment and low-abundance filtering) could lead to biased results, if sparse data (i.e. *Antarctic soils* with low diversity Wall, 2012; Zhan et al., 2014; and low overall sequence count) or an artificial community (i.e. *Artificial blend*, with high overall sequence count, but little sequence diversity) were used for calibration. In the first case (calibration with *Antarctic soils*), the amount of sequence artefacts could be underestimated, when applied to the *Artificial*

blend. In the second case (calibration with *Artificial blend*), the amount of sequence artefacts could be overestimated in data from *Antarctic soils*. Consequently, we chose replicate metagenetic data from a higher diversity and richer natural community (i.e. *Australian soils*) for thresholds adjustment. We considered that these data, would represent a compromise between over- or under-estimating the amount of sequence artefacts. Application of analysis parameters determined with *Australian soils* to *Antarctic soils* resulted in exclusion of a large proportion of phylotypes, which could be obtained when less stringent parameters are used (see Methods and chapter 4); but for the purpose of this study, we favoured well-defined (i.e. tightly clustered, taxonomically closely assigned, highly abundant) 18S and COI phylotypes. While this approach may appear puzzling at first glance, it allowed us to compare high quality metagenetic data to morphological approaches. Subsequently, we were able to show that both metagenetic data generation and morphological taxonomy assignments were much more complicated for small, cryptic Antarctic terrestrial invertebrates than for larger, less cryptic invertebrates of higher diverse and taxon-rich ecosystems where also a comparatively larger reference database exists. Since we conducted chimera filtering and low-abundance filtering, and also excluded non-Antarctic phylotypes from *Antarctic soils*, miss-assignments in metagenetic Antarctic data are less likely to reflect phylotype artefacts or contamination. Thus, our approach also allowed detection of Antarctic phylotypes that are not contained in our reference data, which are likely to have been assigned with incorrect taxonomic information from this incomplete reference data (i.e. insects in *Antarctic soils* for COI, Fig. 3.6c).

3.4.2. Detecting highly abundant and cryptic Antarctic invertebrates

Metagenetic approaches are preferable over morphological biodiversity assessment particularly for the highly abundant and most cryptic Antarctic nematodes and rotifers. Multiple comparisons between metagenetic and morphologic biodiversity assessments have shown that retrieval of a completely overlapping species inventory is difficult to achieve due to inherent biases of each approach (Cowart et al., 2015; Tréguier et al., 2014; Wu et al., 2009). Nematodes are often missed in morphologic approaches due to constraints of extraction methods (Cowart et al., 2015; Wu et al., 2009). Antarctic rotifers are hard to classify on low taxonomic ranks due to their conserved morphology and small size (Dartnall, 1983; Velasco-Castrillón et al., 2014).

Both phyla can occur in high abundances in Antarctic (Sohlenius and Bostroem, 2008; Sohlenius et al., 1996). While metagenetic markers perform differently in detecting expected phylotypes in an DNA mixtures (Clarke et al., 2014), both markers employed here were able to provide family level assignments to nematodes and rotifers. The high abundance of those taxa constitutes a constraint to morphological approaches and increases their DNA contributions to low-diverse Antarctic soil extracts, leading to high success of metagenetic approaches (Cowart et al., 2015; Wu et al., 2009). Apart from Araeolaimida (Nematoda), and Adinetida (Rotifera), our phylotype data did not contain morphotypes detected by the morphological assessment. These absences may be caused by (a) absences of target organisms in the sample, (b) inappropriate DNA extract composition, (c) poor inappropriate reference data (Bik et al., 2012; Cowart et al., 2015; Egan et al., 2015; Tréguier et al., 2014). While we cannot rule out that sub-samples of *Antarctic soils* used for extractions lacked taxa found in morphologic assessments (a), extracting large quantities of soil makes biased DNA extract composition (b) unlikely (Taberlet et al., 2012), particularly since our extraction method had been optimised for the detection of soil invertebrates (Ophel-Keller et al., 2008; Pankhurst et al., 1996). Overall lower amplicon concentrations for COI (Tab. 3.1) indicated lower PCR performance in comparison to 18S (c), but retrieval of invertebrate phylotypes was nonetheless possible (chapter 4). We were aware of the possibility to retrieve phylotype information from tardigrades in (18S) and chelicerates (COI) with the employed markers in addition to the detected phylotypes (chapter 4), but chose not to do so in order to only yield precisely defined invertebrate phylotypes with parameters chosen to exclude low-abundant phylotypes (Fig. 3.3). In consequence incorrect taxonomy assignment to phylotypes also detected among morphotypes is unlikely (d). Our results hence showed that both 18S and COI markers were well suited to accurately detect Antarctic rotifers and nematodes on the family level from bulk soil extracts. Additionally, the employed 18S marker was able to accurately detect Oribatida (Chelicerata), which may have been missed entirely with the morphologic approach.

3.4.3. Metagenetic marker choice for Antarctic invertebrates

Metagenetic markers employed for the detection of Antarctic invertebrates from bulk soil extracts need to be chosen depending on the desired rank-resolution of taxonomic assignments, and with regard to available reference data. The relatively slow mutation

rate of 18S rDNA made it a widely applied marker region to investigate high-rank relationships among metazoans, while the faster mutation rate of the mitochondrial COI region was considered well-suited for the delimitation among lower taxonomic ranks (Abouheif et al., 1998; Medlin et al., 1988; Moritz et al., 1987; Wiemers and Fiedler, 2007). Although metagenetic 18S data may fail to accurately reflect biodiversity in mixed samples, the 18S gene region was considered an efficient and powerful marker for profiling unknown communities (Wu et al., 2011; Zhan et al., 2014). At the same time, while COI was sometimes assumed to perform better in describing lower level diversity in metagenetic data, recent studies showed that this might not be the case, due to high variability in the COI region, constraining the ability to design metagenetic markers (Deagle et al., 2014; Zhan et al., 2014). Our results provided details for which taxonomic ranks both 18S and COI markers might be best suited when describing the biodiversity of Antarctic soil invertebrates: In the *Australian blend*, COI performed better in retrieving morphologically concordant class and order level information, while on the family level 18S yielded higher concordance. Morphological assignments in *Antarctic soils* were constrained by the comparatively small invertebrates contained in these samples (supplemental Fig. 3.8), but for the Antarctic sample (LH-2) with a usable ICC (0.752) closest to the reference value (0.843) COI performed better on phylum and class level, while 18S retrieved better concordance at the order to species ranks. The overall decreased low-rank performance of COI in *Antarctic soils* is likely due to large variability of this marker in concordance with other studies (Berney et al., 2004; Deagle et al., 2014; Kwong et al., 2012). We therefore recommend the application of COI markers for Antarctic invertebrate biodiversity assessments only for high taxonomic ranks, and to complement phylotype information obtained through other markers, such as 18S, as exemplified in Chapter 4. On lower taxonomic ranks, the taxonomic resolution of 18S outperforms the taxonomic resolution of COI in metagenetic approaches for the biodiversity assessment of Antarctic invertebrates.

3.5. Conclusions

We compared the fidelity of taxonomic assignments yielded through morphologic and metagenetic approaches, using two molecular markers offering more substantial reference data than available for other metagenetic markers. We showed that metagenetic approaches employing 18S and COI are particularly well suited in

detecting the smallest and most cryptic Antarctic invertebrates, nematodes and rotifers, in bulk soil extracts. Members of the phylum Chelicerata (Antarctic oribatid mites) may be detected with metagenetic approaches, even when missed in taxonomic assignments. On low taxonomic ranks 18S data currently outperforms COI in accurately recognising Antarctic invertebrates phylotypes due to a lack of reference data for the COI marker.

3.6. Acknowledgements

We thank Alejandro Velasco-Castrillon (University of Adelaide) for contributing Antarctic invertebrate morphotype information, Stephen Pederson and Greg R. Guerin (University of Adelaide) for helpful discussions on analysis coding, Jimmy Breen (University of Adelaide) for maintaining the computational infrastructure required for sequence processing.

3.7. Supporting information

Supporting information is provided subsequently and at <https://zenodo.org/record/19178>¹

¹Data pre-release with closed access. Data will be available publicly after manuscript publication.

3.8. Supplemental information: Matching phylotypes and morphotypes to invertebrate taxonomic assignments: implications for metagenetic surveys in terrestrial Antarctica

3.8.1. Methods and Materials

Provided are details on primer design, deconvolution and chimera screening, clustering, phylotype filtering, determination of Antarctic phylotypes and taxonomic information of phylotypes and morphotypes.

3.8.1.1. Primer design

The forward 18S primer sequence consisted of an Illumina adapter, a primer pad and linker, as well as the target-priming region “1391f” (Gilbert et al., 2010; Parfrey et al., 2014). The reverse primer sequence contained the reverse complement of a 3' Illumina adapter, a twelve bp recognition sequence allowing error correction (Caporaso et al., 2012; Golay, 1949) (assigned to samples as listed Tab.3.2), a reverse primer pad and linker, as well as the reverse primer “EukBr”(Gilbert et al., 2010). Samples discussed here were processed as part of a sequencing project consisting of 192 samples. Only 140 adequate recognition sequences are readily available and tested (Faircloth and Glenn, 2012) for multiplexing on the 454 platform (Roche, 2009a). Some combinations of adapter, recognition and primer sequences may inhibit amplification (Vallone and Butler, 2004). The selection of 454 primers used here featured only sequences that did not exhibit hairpins (tested at <http://www.thermoscientificbio.com/webtools/multipleprimer/>). The reduced amount of possible primer sequences required sequencing of amplicons in two orientations to allow deconvolution (see Tab.3.2). Sequence constructs of the first sequencing orientation consisted of a primer with 454 adaptor “Lib-L Primer A key” (Roche, 2009a,b), recognition sequences unique to sample (Tab.3.2) as well as the primer “mlCOIintF” (Leray et al., 2013); furthermore of a primer linking the reverse complemented 454 adaptor “Lib L Primer B key” (Roche, 2009a,b) with the reverse complemented sequence “HCO2198” (Folmer et al., 1994). Sequence constructs of reversed sequencing orientation contained the 454 adaptor “Lib-L Primer A key” (Roche, 2009a,b), recognition sequence unique to sample (Tab.3.2)

as well as primer “HCO2198” in the first sequencing oligonucleotide. The second oligonucleotide contained the reverse complemented 454 adaptor “Lib L Primer B key”, and “mlCOIintF” (Leray et al., 2013). Further details are provided in sec. 2.8.

3.8.1.2. Amplification and sequencing

Triplicate PCRs were prepared containing 2 µl template, 1.5 mM MgCl₂, 1 x AmpliTaq Gold buffer (Thermo Fisher Scientific, Waltham, US-MA), 0.25 mM dNTPs, 0.5 µM forward and reverse primer and 1.25 units AmpliTaq Gold (Thermo Fisher Scientific, Waltham, US-MA) in a 20 µl reaction. Thermal cycling conditions were set to initial denaturation at 94 °C (3 min), followed by 35 cycles (94 °C for 0:45 min, 57 °C for 1 min, and 72 °C for 1:30 min) with a final elongation of 10 minutes (72 °C). Long extension times were used to counteract chimera formation (Lenz and Becker, 2008; Yu et al., 2012). Amplicons were visualised on 2% agarose gels.

Combined and un-quantified replicate amplicons of each gene region were purified using Agencourt AMPure XP (Beckman Coulter, Brea, US-CA), quantified using Qubit QuantiFluor dsDNA kits (Promega, Fitchburg, US-WI). Amplicons above 0.25 ng/µl (Tab. 3.1) were then pooled by weight (18S: 12 ng/µl per sample, COI: 4 ng/µl per sample; except first replicate of each of the two *Australian soils*, COI: 3 ng/µl per sample, due to overall lower amplicon concentration). Concentration in library pools was established using a 2100 Bioanalyser (Agilent Technologies, Santa Clara, US-CA) and diluted to 9 pM for Illumina sequencing (18S) or concentrated to 3.18 ng/µl for emulsion PCR preceding 454 sequencing (COI). Libraries were paired-end sequenced in two separate runs on the Illumina MiSeq platform (Illumina, San Diego, US-CA) with 300 cycles or on two separate quarter of a 454 PicoTiterPlate (454 GS FLX platform).

DNA extraction and PCR controls were included into amplification and sequencing for both genes if the cleaned amplicons were sufficiently concentrated (required were 0.25 ng/µl). Among 192 combined PCR reactions for 18S, six of 13 PCR controls and one of three blended extraction controls were sequenced. For COI, seven of 13 PCR controls and two of three blended extraction controls were sequenced. Usable sequence data for contamination removal could be obtained for 18S, COI data contained no invertebrate phylotypes that would have interfered with subsequent analysis steps.

3.8.1.3. Reference data for taxonomic assignments

Unclustered rather than clustered reference data was chosen for 18S and COI to allow more precise taxonomic assignment of phylotypes using QIIME 1.8 (Caporaso et al., 2010). For 18S, SILVA reference data (Pruesse et al., 2007) release 111 was used, consisting of 20,115 metazoan entries (9,904 arthropods, 1,941 nematodes, 126 rotifers, 197 tardigrades). Reference data for COI was compiled from earlier Antarctic studies (Velasco-Castrillón et al., 2014) as well as GenBank (Benson et al., 2011). Antarctic records for COI consisted of 42 arthropods and 621 nematodes, rotifers and tardigrades. GenBank COI records were retrieved on the 16 January 2015, and consisted of sequences between 100 – 1000 bp length, with 651,565 arthropods, 4,102 nematodes, 3,724 rotifers and 598 tardigrades. Combined reference data was de-replicated using 100% clustering with UCLUST (Edgar, 2010). The final COI reference collection included 79,906 invertebrate sequences (74,603 arthropods, 2,794 nematodes, 1,990 rotifers, 519 tardigrades).

3.8.1.4. Deconvolution and chimera screening

All analysis steps were performed with scripts provided by QIIME 1.8. (Caporaso et al., 2010). Deconvolution of 18S data was accomplished using QIIME's "split_libraries.py" with parameters "-q 10" (maximum unacceptable Phred score) "--store_qual_scores" (quality scores not discarded), "--rev_comp _mapping_barcodes" (reverse-complementing sequences before writing to output file). COI data was deconvoluted after reverse-complementing sequence data with opposite orientations (Tab.3.2) using QIIME's "split_libraries.py" with parameters "-l 350" (minimum sequence length), "-L 400" (maximum sequence length), "-b 12" (enabling error correction of Golay barcode), "-M 3" (allowed maximum primer mismatches) "-w 50" (sliding window test of quality scores, default), "-z truncate_only" (truncation of reverse primers) and "-d" (recording quality scores). The effect of chimeras in metagenetic data is detrimental (Bokulich et al., 2013; Edgar, 2013). Chimera screening of unclustered 18S data using USEARCH (Edgar, 2010) was not performed due to software licensing restrictions², in unclustered COI data 37 *de novo* chimeras were detected and removed. In an alternative approach, to reduce the effect of chimeras, phylotypes with abundances <5 sequences across the 18S or COI data were removed in later analysis steps (Carew et al., 2013). De-nosing

²No licence-free processing of files larger than 5 GB

of COI 454 data was omitted, after initial test showed this to only decreased the phylotype count by 0.2 % while drastically lowering the quality of correctly called phylotypes: in de-noised 454 data, initial flowgram clustering resulted in spreading of high quality sequences with imperfectly trimmed adapter sequences across lower quality sequence data and resulted in clustering bias based on amplicon orientation. Without denoising this effect was circumvented. Additionally, initial analyses showed that adapter trimming, even with consecutive application of several tools such as Trimmomatic (Lindgreen, 2012) and AdapterRemoval (Lindgreen, 2012), in no case allowed complete removal of adapter sequences from target sequences, meaning that denoising of 454 data by means of flowgram clustering would have had a detrimental effects on overall sequence quality.

3.8.1.5. Clustering

Clustering of 18S data was performed with QIIME “`pick_otus.py`” with default parameters for reasons of computational efficiency. For COI additional parameters were used: “`-optimal`” (every was seed aligned to query), “`--exact_uclust`” (find a match if one exists, even if it is not the best), “`-z`” (reverse strand matching). Further information can be obtained from the UCLUST manual.

3.8.1.6. Phylotype filtering

Phylotype data were filtered in a step-wise process using scripts provided by QIIME as outlined in the online documentation. Specifically, unassigned phylotypes were removed, contamination stemming from control reactions was subtracted (18S only, as for COI no sequence information could be retrieved from extraction and PCR controls) and phylotypes with <5 sequences across data sets were discarded (Carew et al., 2013). After these filtering steps, phylotype data was subset to target organisms that were contained in both reference data sets and also expected in the Australian / Antarctic samples; arthropods, nematodes, tardigrades and rotifers. To mitigate the effect of spurious taxonomy assignments during later analysis steps all phylotype tables ($n = 2 \times 7$ for COI and $n = 2 \times 3$ for 18S) were subsequently filtered for low abundant phylotypes with increasing stringency, i.e. from each dataset the lowest 0.1%, 0.2%, 0.3% and 0.5% of phylotypes were discarded. See Tab. 3.3 for filtering results.

3.8.1.7. Determination of Antarctic phylotypes

Origin verification of Antarctic phylotypes was performed using R (R Development Core Team, 2011). Briefly, phylotypes were only considered to be of Antarctic origin, when the Antarctic sequence clusters (i.e. phylotypes) were not also contained in *Australian soils* (also see source code, Appendix B).

3.8.1.8. Taxonomic information of phylotypes and morphotypes

Comparison between phylotypes and morphotypes of corresponding samples was only possible, when sequence data for 18S or COI phylotype data was available for morphologically assessed samples. Initially, morphotype information for Antarctic samples was converted into file formats accessible through QIIME and Phyloseq. In R, NCBI taxonomy information was used to correct taxonomy information carried through from 18S and COI reference databases to match a format used in the morphotypes data, the taxonomic ranks phylum, class, order, family, genus and species were chosen (also see source code, Appendix B).

3.8.2. Tables and Figures

Table 3.1.: Sample collection locations and molecular lab work details. Concentration measurements are provided after pooling of three PCR replicates and small fragment removal (*). Required for handling were 0.25 ng / µl (**). Concentrations are given for both replicates of Australian soil controls (***).

Origin	Name	Lat (°S)	Long (°E)	Location	Amplified	18S PCR [ng/µl] *	COI PCR [ng/µl] *
Antarctic soil	LH-1	-69.37552	76.38286	Larsemann Hills	18S	6	0.58
	LH-2	-69.40918	76.00242	Larsemann Hills	18S, COI	6	1.72
	HI-1	-68.82586	77.72085	Hop Island	18S	6	1.16
	VH-1	-68.6415	78.29643	Vestfold Hills	18S	6	1.34
	VH-2	-68.60162	78.34634	Vestfold Hills	18S	6	1.04
	CS-1	-66.28059	110.52003	Casey Station	18S, COI	6	0.14 **
	CS-2	-66.28297	110.52717	Casey Station	18S, COI	5.65	0.26
	Soil 1	-34.96884	138.63684	Adelaide University	18S, COI	6,6 ***	1.35,1.74 ***
Australian soil	Soil 2	-34.96880	138.63687	Adelaide University	18S, COI	6,6 ***	1.48,1.64 ***
	Australian blend	-34.99983	138.79951	Adelaide Hills	18S, COI	6	1.04

Table 3.2.: Recognition sequences enabling assignment of 18S and COI sequence data to samples. One sample did not meet the required minimum concentration for sequencing of 0.25 ng / μ l (*). Double use of sequencing tag possible since two separate 454 runs were conducted.

Origin	Sample	18S sequence	COI sequence	Orientation of COI amplicon
Antarctic soil	LH-1	TGTAACGCCGAT	ACGCGATCGATA	rev
	LH-2	TAACGTGTGTGC	CAGTAGACGTGG	fwd
	HI-1	CAGCTCATCAGC	TACGCTGTCTGG	fwd
	VH-1	AGCAGAACATCT	- (*)	-
	VH-2	GCAACACCATCC	TCTAGCGACTGG	fwd
	CS-1	TGGAGTAGGTGG	AGCTCACGTAT	rev
	CS-2	ATGTCACCGCTG	TGACGTATGTTA	rev
	Australian soil	Soil 1	TTGACGACATCG	rev
			TGTGCGATAACA	rev
		Soil 2	ACATACTGAGCA	rev
			GATTATCGACGA	rev
			GATCCCACGTAC	rev
Australian blend			AGTGCTACGATA	rev

Table 3.3.: Decreasing phylotype counts for 18S and COI data sets during filtering. Given are phylotype and sequence counts (“- / -”). Values are only given for final analysis parameters. No phylotypes were retrieved for COI PCR and extraction controls.

Gene	Sample	All	Assigned	w/o blanks	Invertebrates <5 sequences	Abundance filtering	Antarctic
18S	LH-1	2319 / 120285	2003 / 115152	1926 / 73647	13 / 120	8 / 95	4 / 54
	LH-2	3196 / 156486	2821 / 150038	2699 / 54539	9 / 748	5 / 733	0 / 0
	HI-1	2287 / 132684	1656 / 103127	1598 / 82937	16 / 8436	3 / 8419	1 / 8295
	VH-1	2830 / 129315	2636 / 123652	2568 / 19744	51 / 273	4 / 191	0 / 0
	VH-2	1992 / 130811	1920 / 130412	1877 / 4403	3 / 10	2 / 9	0 / 0
	CS-1	1776 / 78544	1577 / 76673	1507 / 57437	2 / 9	2 / 9	0 / 0
	CS-2	1285 / 98865	1190 / 94889	1147 / 92395	7 / 466	4 / 457	2 / 453
	Soil 1	9318 / 139793	7844 / 129229	7621 / 61682	110 / 734	11 / 306.0	- / -
		7162 / 133522	5627 / 122301	5437 / 57965	97 / 767	13 / 352.0	- / -
	Soil 2	8990 / 166642	7519 / 157171	7303 / 83038	103 / 1945	11 / 306.0	- / -
		6492 / 129474	5019 / 120176	4868 / 63529	91 / 1781	13 / 352.0	- / -
	Australian blend	1186 / 94231	1127 / 93352	1098 / 93123	45 / 91768	13 / 90829	- / -
COI	LH-1	28 / 507	15 / 367.0	- / -	9 / 356	4 / 340	0 / 0
	LH-2	50 / 2274	18 / 750	- / -	17 / 749	9 / 692	2 / 11
	VH-1	- / -	- / -	- / -	- / -	- / -	0 / 0
	VH-2	17 / 2434	2 / 106	- / -	6 / 106	4 / 101	0 / 0
	HI-1	64 / 1431	27 / 538	- / -	18 / 514	6 / 424	0 / 0
	CS-1	36 / 619	13 / 367	- / -	9 / 359	4 / 335	1 / 41
	CS-2	24 / 552	12 / 308	- / -	10 / 305	6 / 279	1 / 15
	Soil 1	449 / 1874	204 / 666	- / -	76 / 501	12 / 232	- / -
		297 / 917	131 / 328	- / -	63 / 245	11 / 112	- / -
	Soil 2	416 / 1898	194 / 820	- / -	73 / 660	13 / 413	- / -
		237 / 836	100 / 352	- / -	52 / 295	12 / 190	- / -
	Australian blend	19 / 788	18 / 787	- / -	10 / 770	6 / 725	- / -

Table 3.4.: Intraclass correlation coefficients (ICC) of taxonomic assignments using 18S and COI markers when compared to morphologic reference information. A value of 1.0 indicates total concordance between metagenetic and morphologic taxonomic assignments.

Australian blend	CS -1	CS-2	HI-1	LH-2
0.843	0.429	1.000	-0.173	0.759

3.8.3. Analysis code

Analysis code is provided in Appendix B.

Species		Families	
<i>Drymaplaneta communis</i>			Acrididae
			Blattidae
			Chrysopidae
		Ischnura	Coenagrionidae
			Eurybrachyidae
			Forficulidae
			Formicidae
			Ichneumonidae
			Lauxaniidae
			Lymantridae
			Rhinotermitidae
			Scarabaeidae
			Sparassidae
			Tettigoniidae
			undetermined

Figure 3.7.: Reference data composition of *Australian blend*. Invertebrate composition is shown on family (horizontal facets), genus (colours) and species level (columns). Reference data is completely defined on the family level, genus-level information is available for many taxa (colours) and species information is available for three taxa.

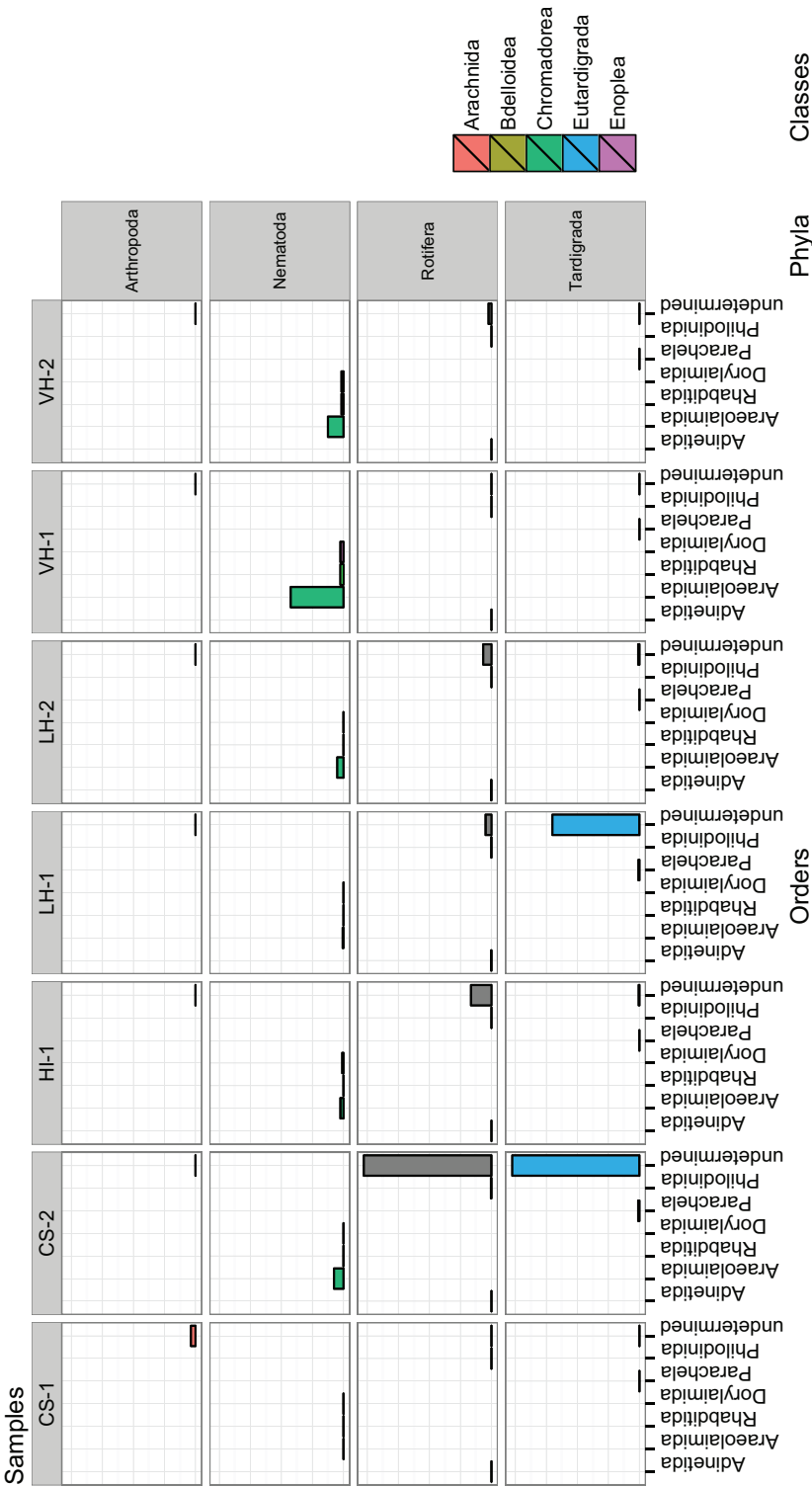



Figure 3.8.: Reference data composition of *Antarctic soils*. Composition of the seven *Antarctic soils* (in vertical facets) is shown by phylum (horizontal facets), class (colours) and order levels (columns). Reference data is completely defined on the phylum level, class information is available for many taxa (colours) and order information is available for six taxa.

Statement of Authorship

Title of Paper	The terrestrial invertebrates of the Prince-Charles Mountains, East Antarctica: Distribution in relation to soil nutrients and substrate composition.
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Publication Style
Publication Details	We examined the invertebrate biodiversity from 103 bulk soil extracts from the Prince Charles Mountains, Antarctica in conjunction with the mineral and soil geochemical composition. Correlations between the distribution of biodiversity and abiotic data indicate that niches occupied by soil invertebrates may be explained, in part, by soil nutrient content, mineral composition, but also by remoteness (e.g. latitude).


Principal Author


Name of Principal Author (Candidate)	Paul Czechowski		
Contribution to the Paper	Participated in field work, handled samples during extraction and sub-sampling. Arranged soil geochemical analysis and logistics for mineral analysis. Conducted lab work for mineral analysis. Designed and applied experiments and analysis approaches, interpreted results, designed and structured draft manuscripts, wrote manuscript, designed and created figures and tables.		
Overall percentage (%)	80%		
Signature		Date	9.6.2015


Co-Author Contributions

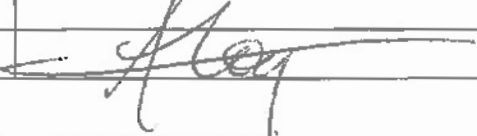
By signing the Statement of Authorship, each author certifies that:

- the candidate's stated contribution to the publication is accurate (as detailed above);
- permission is granted for the candidate to include the publication in the thesis; and
- the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Stevens		
Contribution to the Paper	Retrieved funding for, organized, planned, coordinated field campaign in Antarctica. Contributed ideas on experiment and analysis design, interpretation of results, and manuscript structure, revised and edited draft and main manuscripts, contributed ideas for figures and tables. Facilitated access to laboratory facilities, equipment and reagents necessary for soil analysis.		
Signature		Date	11.6.2015

Name of Co-Author	Duanne White		
Contribution to the Paper	Provided infrastructure and hospitality for mineral analysis. Processed raw mineral data for inclusion into downstream analyses. Provided information and comments regarding the mineral composition of the sampling area.		
Signature		Date	11.6.2015

Name of Co-Author	Laurence Clarke		
Contribution to the Paper	Contributed ideas on experiment and analysis design, interpretation of results, and manuscript structure, revised and edited draft and main manuscripts, contributed ideas for figures and tables.		
Signature		Date	15.6.2015

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Contributed ideas on experiment and analysis design. Provided access to laboratory facilities, equipment and reagents necessary for molecular analysis. Provided and facilitated access to computational infrastructure for data analysis. Edited main manuscript.		
Signature		Date	12.6.2015

Please cut and paste additional co-author panels here as required.

4. Salinity gradients determine invertebrate distribution in the Prince Charles Mountains, East Antarctica

Paul Czechowski¹, Duanne White², Laurence J. Clarke^{1, 3, 4}, Alan Cooper¹, Mark I. Stevens^{5, 6}

¹ Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005 Australia; ² Institute for Applied Ecology, University of Canberra, Canberra, ACT 2601 Australia; ³ Australian Antarctic Division, Channel Highway, Kingston, TAS 7050, Australia; ⁴ Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, TAS 7001, Australia; ⁵ South Australian Museum, GPO Box 234, Adelaide SA 5000, Australia; ⁶ School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, SA 5000, Australia

Abstract Knowledge of the relationship between terrestrial Antarctic biodiversity and environmental constraints is necessary to understand the potential effects of human-mediated environmental change and for successful conservation management. A valuable approach to explore the potential impact of environmental change on ecosystems is to compare biodiversity patterns across large-scale environmental gradients. However, studies to date have been limited by the cryptic nature of many Antarctic invertebrate taxa and the time-consuming nature of morphology-based biodiversity assessments. We used *high throughput sequencing* (HTS)-derived meta-genetic biodiversity information to elucidate the relationship between soil properties and invertebrate biodiversity in the Prince Charles Mountains, East Antarctica. We analysed data obtained from 103 soil samples collected across three broad sampling

regions (Mount Menzies, Mawson Escarpment, and Lake Terrasovoje). We found invertebrate distribution in the Prince Charles Mountains was significantly influenced by electrical conductivity and/or sulphur content and, to a lesser extent, slope and elevation. The classes Enoplea (Nematoda), Bdelloidea (Rotifera), and phyla Tardigrada and Arachnida, only occurred in low salinity substrates with relatively abundant nutrients, but Chromadorea (Nematoda) and Monogonata (Rotifera) were apparently less influenced by salinity because they showed broader distributions. Positive correlation between soil salt concentration and terrain age (time since deglaciation) indicated that terrain age may indirectly influence Antarctic terrestrial biodiversity. We demonstrate the value of metagenetic HTS approaches to investigate Antarctic environmental constraints on the distribution of terrestrial invertebrates across large spatial scales.

Keywords Antarctica, invertebrates, environmental DNA, gradient, salinity, high throughput sequencing (HTS), Prince Charles Mountains

4.1. Introduction

Biodiversity information and its relation to environmental constraints is required to address the effect of anticipated human-mediated environmental change on Antarctic terrestrial life (Kennicutt et al., 2015), and is required for successful conservation management (Chown et al., 2012; Terauds et al., 2012). A valuable approach to explore the potential impact of environmental change on ecosystems is to compare biodiversity patterns across environmental gradients (Howard-Williams et al., 2010). For example, comparing ecosystems at slightly warmer or slightly cooler sites, e.g. with latitude or altitude, may allow future predictions of biodiversity changes in response to increasing temperature (Howard-Williams et al., 2006). However, a limited number of sites may prevent conclusions about latitudinal and/or climate controls over patterns of biodiversity (Barrett et al., 2006). Baseline data enabling predictions of future environmental changes across Antarctica should hence describe biodiversity with broad taxonomic focus across large spatial scales in relation to as many potentially environmental variables as possible (Convey et al., 2014; Gutt et al., 2012).

It has previously been suggested that geo-glaciological events and the presence of past refuges may be more important than latitudinal variations in climatic and environmental conditions in determining the large-scale distributions of most Ant-

arctic terrestrial fauna (Caruso et al., 2010). On smaller spatial scales invertebrate biodiversity was frequently associated with low salinity and high nutrient content (Magalhaes et al., 2012; Powers et al., 1998; Velasco-Castrillón et al., 2014), with the exception of nematodes (Freckman and Virginia, 1997) and rotifers (Barrett et al., 2006). Typically, a set of interrelated soil and environmental factors determined the abundance and composition of Antarctic soil communities (Courtright et al., 2001) and hence eigenvector methods were found well-suited to study such relationships (Caruso et al., 2010). A study linking many environmental variables to all major Antarctic invertebrates using an eigenvector-based approach thus may be able to elucidate whether the distribution of Antarctic invertebrate taxa on large spatial scales is strongly influenced by past geo-glaciological events or rather correlated with environmental constraints in a similar fashion as observed on small spatial scales.

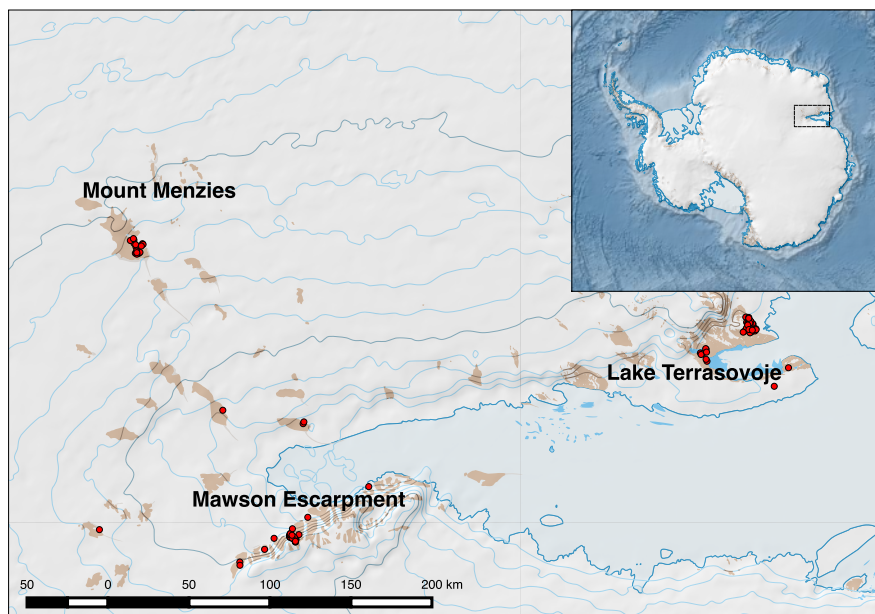


Figure 4.1.: Sampling locations in the Prince Charles Mountains, East Antarctica. Invertebrate phylotype and observation metadata was obtained from 103 samples and analysed here. Phylotype information was combined from 47 samples using 18S and COI markers, and from 56 samples exclusively using 18S. Analysed environmental metadata included 21 variables describing substrate geochemical and mineral properties across 141 samples; from 103 of these biological data could be obtained.

High throughput sequencing (HTS) is a promising approach to rapidly obtain biodiversity information from a large number of samples in extreme environments, such

as Antarctica (Cowan et al., 2015). For Antarctic invertebrates, morphological approaches require a high level of taxon-specific knowledge and are practically limited to small sample numbers (Colesie et al., 2014; Sohlenius and Bostroem, 2008; Sohlenius et al., 1996; Zawierucha et al., 2015). Additionally, morphologically cryptic Antarctic species may consist of multiple genetic lineages shaped by long-term isolation, making molecular approaches a more suitable tool to investigate their diversity (Convey and Stevens, 2007; Rogers, 2007). Early Antarctic metagenetic studies used bacterial cloning to investigate drivers of Antarctic biodiversity across large-scale spatial gradients (Convey et al., 2014; Lawley et al., 2004). These cloning approaches were work-intensive (Fell et al., 2006; Lawley et al., 2004; Nakai et al., 2012), and more recent high-throughput approaches are better suited to generating biodiversity information from large sample numbers (reviewed in Bik et al., 2012; Bohmann et al., 2014). HTS-based metagenetic approaches have now been used to describe invertebrate distribution and diversity on a global scale (Wu et al., 2011). These methods could also provide valuable information regarding Antarctic invertebrate biodiversity under the influence of environmental variables (Chown et al., 2015).

In this paper, we used HTS-derived metagenetic biodiversity information to elucidate the relationship between soil properties and invertebrate biodiversity in the Prince Charles Mountains (PCMs), East Antarctica. We analysed data obtained from a total of 103 soil samples collected across three broad sampling regions (Mount Menzies, Mawson Escarpment, and Lake Terrasovoje). Initially, we characterized the spatial variation of soil geochemical and mineral composition, as well as the distribution of the four major Antarctic invertebrate (sub-) phyla (nematodes, rotifers, tardigrades and chelicerates) among all samples, on class level. Using constrained ordination of environmental and biological observations we then identified potential environmental variables influencing invertebrate diversity and distribution. We show that the distribution of invertebrates in the Prince Charles Mountains is mainly influenced by electrical conductivity, sulphur content, slope and elevation, suggesting that long-term soil formation processes such as age-related salt accumulation unique to Antarctica are driving invertebrate distribution across large spatial scales. This effect is less pronounced at coastal sites, where substrate salinity is relatively low, but nutrient input relatively high.

4.2. Methods

4.2.1. Fieldwork

Fieldwork was conducted in the Prince Charles Mountains (East Antarctica) between 26th November 2011 and 21th January 2012 at Mount Menzies, Mawson Escarpment and Lake Terrasovoje. Satellite imagery was used in ARC-GIS v10.0 (Esri, Redlands, US-CA) to determine sampling locations across the three regions based on broader glaciological and geological properties (bedrock, moraine lines and altitude) (Fig. 4.1). The centroid of each sampling location was used as the sampling site. At each site, a maximum of 500 g of soil was collected from the top 10 cm of the substratum by combining five subsamples from the corners and centre of a one metre square quadrat into a sterile WhirlPak bag (Nasco, Fort Atkinson, US-WI; protocol after Magalhaes et al. 2012; Velasco-Castrillón et al. 2014). Sample contamination was minimised by wearing nitrile gloves and cleaning equipment with 70% ethanol. Additionally, at each sampling site a maximum of 300g of soil was collected in the same fashion for soil geochemical and mineral analysis. A total of 103 soil samples were obtained and all samples were stored at -30 to +4 °C in the field in insulated containers (Coleman, Wichita, US-KS). Samples were transported and stored at -20 °C.

4.2.2. Soil geochemical and mineral analysis

Soil geochemical analysis was performed at the CSBP Soil and Plant Analysis Laboratory (Bibra Lake, AU-QLD) to determine colour, texture, electrical conductivity, pH (for CaCl₂ and H₂O), gravel content, NH₄⁺, NO₃⁻, P, K, S, and organic C for each sample. Geochemical analysis methods, sourced from Rayment and Lyons (2011), are listed in Tab. 4.1. Electrical conductivity was used as a proxy for salinity (Magalhaes et al., 2012). For the collection of X-ray diffraction spectra, all soils were dried at 100 °C for 48 h and then each consecutively sieved through 2 mm and 63 µm meshes. X-ray diffraction spectra of the resulting powders were measured using a BTXII Benchtop XRD (Cu-Kα X-ray source), with 105 consecutive cycles per sample. Mineral identification was conducted using PANalytical's Highscore Plus software v3.0e, against the open crystallographic database (Grazulis et al., 2012). Mineral groups were considered present if position and intensity of phase-identified peaks matched three or more peaks in the database. Semi-quantitative measures of mineral abundance were determined as described in Chung (1974) for quartz,

feldspar, titanite, pyroxene / amphibole / garnet, micas, dolomite, kaolin / chlorite, calcite and chlorite.

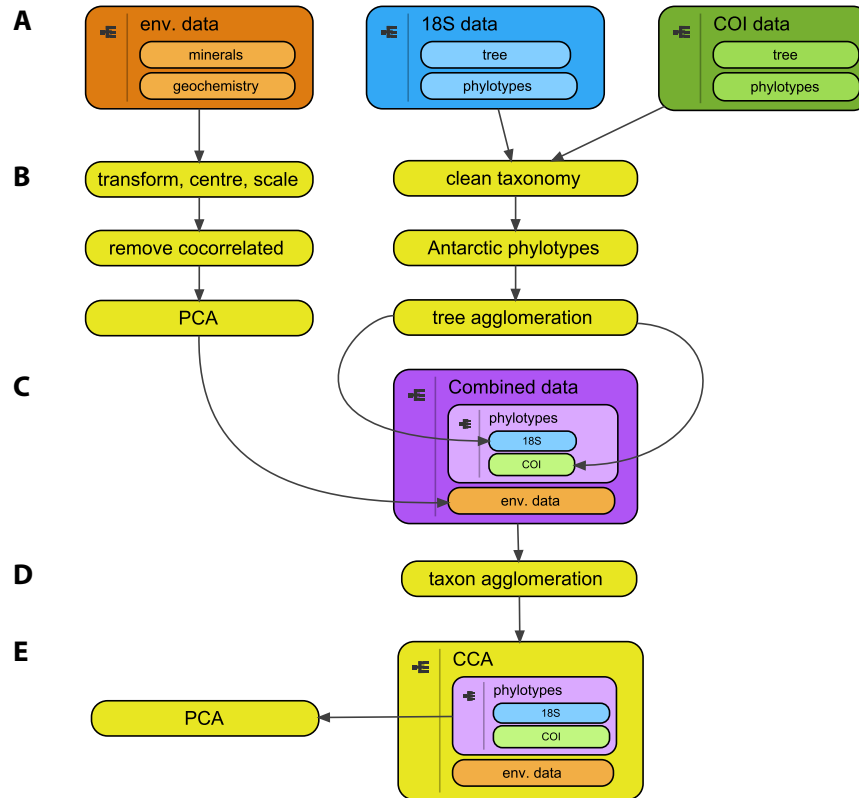


Figure 4.2.: Methods performed to analyse metagenetic invertebrate data in conjunction with geochemical and mineral observations. (A): Geochemical and mineral observations (orange) were imported into R, along with QIIME-generated invertebrate phylotype information and matching phylogenetic invertebrate reference trees for 18S and COI (blue and green, respectively). (B): Environmental data was analysed using Principal Component Analysis (PCA) before and after transformation, removal of co-correlated variables and removal of outliers, PCA on pre-processed data was used for analysis. 18S and COI invertebrate phylotype taxonomy was determined to superphylum, phylum, class, order, genus and species level (where possible) using NCBI taxonomy information. Decreased site heterogeneity was required to enable *Canonical Correspondence Analysis* (CCA); taxon-agnostic tree clipping of the 18S and COI reference trees was therefore used to reduce phylotype numbers. (C): Biological data sets and environmental data were combined. (D): Agglomerating taxa on the class level further reduced site heterogeneity among biological data. (E): To relate biological and observation metadata, an empty CCA model was defined, and each of the 21 observation variables were added step-wise towards a fully constrained model. The model with highest-ranking Akaike Information Criterion (AIC) was chosen for interpretation. PCA was used on biological data to complement the analogous analysis of environmental data.

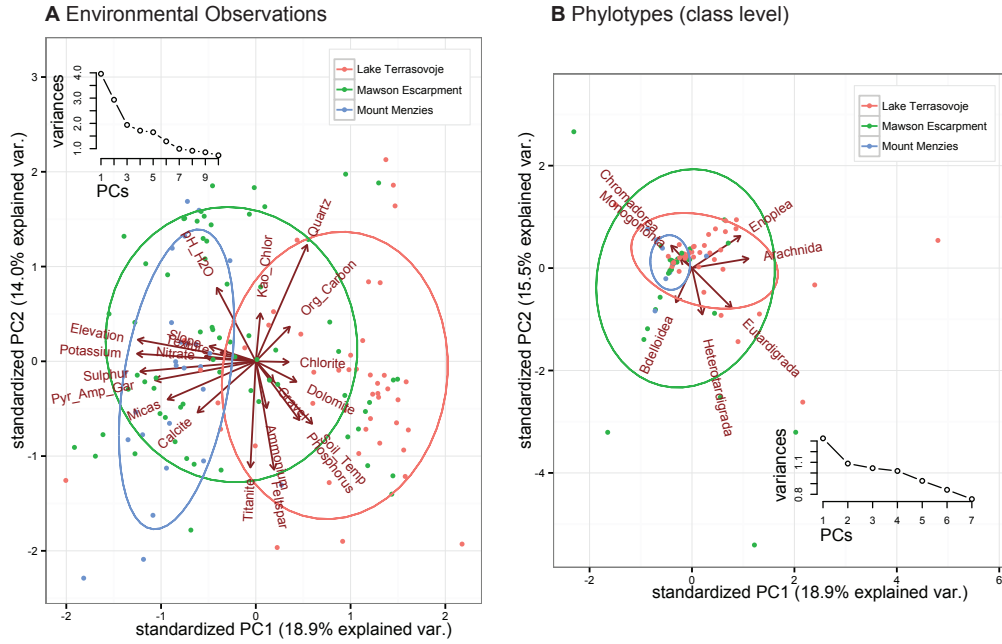


Figure 4.3.: Principal Component Analysis (PCA) of environmental (A) and invertebrate phylotype data (B) for soil samples from Prince Charles Mountains, East Antarctica. Coloured circles indicate normal range of principal components (PCs) for each location. PCA was calculated after pre-processing environmental variables and class-agglomerated invertebrate phylotype counts. Co-correlated variables and outliers were also removed from environmental data (electric conductivity / pH CaCl_2). Insets show variances of first 10 PCs. In 21 environmental principal components 16 contained 95% of all variance, in biological data all PCs were needed to account for 95% variance.

4.2.3. Preparation and analysis of environmental observations

Analyses of biological, geochemical and mineral data were conducted in R v3.2.0 (R Development Core Team, 2011). An overview of the analysis is provided in Fig. 4.2. Principal Component Analysis (PCA) on combined and separated (mineral / geochemical) data was used to compare environmental observations between the 103 samples across 21 variables (and 9 and 9 variables, respectively - without slope, elevation and soil temperature) and plotted using package GGPlot2 v1.0.1 (Wickham, 2009). To meet assumptions of normality and variance uniformity, all observations were Box-Cox-transformed, scaled to unit variance, and centred, with highly (>0.7) co-correlated variables removed (conductivity in favour of S, and pH CaCl_2 in favour of pH H_2O , respectively) (Reid and Spencer, 2009; Sakia, 1992),

using package CARET v6.0-41 (Kuhn, 2008). Since extreme values impeded PCA initially, values with the largest difference from the mean were replaced by their means for all variables (Reid and Spencer, 2009). Geomorphic mapping and weathering studies (White, 2007; White and Hermichen, 2007) and cosmogenic exposure dating (White et al., 2011) were used for age determination of glacial sediment.

4.2.4. Preparation and analysis of biological observations

Methods and materials for molecular laboratory work (DNA extractions, library generation and sequencing) and invertebrate phylotype generation are provided in the supplemental information (sec.4.10) and elsewhere (sec.3.8). Preparation of Antarctic invertebrate phylotype data from COI and 18S metagenetic data was conducted using QIIME v1.8 (Caporaso et al., 2010) and R package PHYLOSEQ v1.12.1 (McMurdie and Holmes, 2013). Reference trees for 18S and COI invertebrate phylotypes were calculated using FASTTREE (Price et al., 2009), via GENEIOUS R7 (<http://www.geneious.com/>), with alignments pre-processed using QIIME default parameters. Sparse biological observations in both data sets resulted in high site heterogeneity, which initially prevented subsequent analysis steps. We used the tree tip agglomeration function (`'tipglom'`) of PHYLOSEQ and reference trees to decrease biological site heterogeneity in a taxon-agnostic fashion (cut height = 0.1). 18S and COI PHYLOSEQ objects were then combined using the PHYLOSEQ functions (`'merge_phyloseq'` and `'merge_phyloseq_pair'`), with taxon and sample data components of each object modified to reflect the origin of the source data (18S or COI). Subsequently, phylotypes were agglomerated to the class level. To relate invertebrate class occurrences to sample origin, class-agglomerated phylotype counts were pre-processed and subjected to PCA as for the environmental observations, but without removal of outliers. In PCAs of higher-rank-agglomerated phylotypes (e.g. class) the effects of possibly biased mixed template sequence abundances (Kanagawa, 2003) are mitigated through the large sample size and combination of multiple phylotypes.

4.2.5. Constrained ordination of environmental and biological observations

To relate environmental observations and invertebrate phylotypes, *Canonical Correspondence Analysis* (CCA) (Ter Braak, 1986) was performed using the R package VEGAN v2.2-1 (Dixon, 2003; Oksanen et al., 2015). CCA models were defined by sequentially adding environmental variables (i.e. constraints) to an initially unconstrained model towards a fully constrained model containing all 21 variables. From these models, the highest-ranking one was selected based on Akaike Information Criterion (AIC) (Bozdogan, 1987). Model evaluation was performed using VEGAN-specific ANOVA functions (all with 1000 permutations), including testing the significance of the final model, its axes, each variable during sequential addition to the model (Type I test) and each variable during sequential elimination from the model (Type III test). Also, variance inflation factors (VIFs) for each variable contained in the final model were obtained.

4.3. Results

4.3.1. Environmental data

When comparing each region to the mean values obtained from all samples (Fig. 4.3a and supplemental Fig. 4.9), the majority of Mount Menzies soils were rich in S, K and quartz. Soils of the Mawson Escarpment encompassed a large variety of environmental properties reflective of the large sampling range, but were comparatively alkaline and K-rich and dominated by micas, calcite and pyroxene, amphibole and garnet. Most soils of Lake Terrasovoje were rich in P, with high gravel content, and relatively high abundance of dolomite and chlorite. Both Mawson Escarpment and Lake Terrasovoje soils had higher levels of NH_4^+ , P and organic C when compared to Mount Menzies. Salt concentrations (conductivity, S, NO_3^- , and P) in soils in the drier inland areas increased with terrain age. The trend was most pronounced at the Mawson Escarpment, where age-salt regressions had R^2 values of 17 to 32%. At the more coastal and relatively humid Lake Terrasovoje sites, correlations were lower (R^2 of 0.1 to 9%), suggesting soluble ions were washed out of the soils by snow- or glacial melt-water rather than accumulating over time. The final PCA included 21 environmental variables (gravel, texture, NH_4^+ , NO_3^- , P, K, S, organic

C, pH H₂O, quartz, feldspar, titanite, pyroxene / amphibole / garnet, micas, dolomite, kaolin / chlorite, calcite, and chlorite, elevation, slope and soil temperature) with the first two principal components (PCs) (Fig. 4.3a) explaining 32.9% of the total variation. Since variance decline was shallow in all PCAs of environmental data (insets Fig. 4.3a and supplemental Fig. 4.9a,b), we focussed on the largest factor loadings of the first two PCs in all analyses.

4.3.2. Biological data

When comparing relative phylotype richness values of each region to the mean values obtained from all samples (Fig. 4.3 and supplemental Fig. 4.12, Fig. 4.13) low abundances of the classes chromadorea (Nematoda) and monogonata (Rotifera) were common in the Mount Menzies area, while abundances were highly variable for all seven invertebrate classes in the Mawson Escarpment region. Most of the Lake Terrasovoje soils exhibited high abundances of classes Enoplea (Nematoda), Arachnids (Chelicerata) and Eutardigrades (Tardigrada).

4.3.3. Biological data in relation to environment

The CCA model with the lowest AIC included 15 variables. Sulphur (S) (and / or co-correlated variable electrical conductivity - see methods), slope and elevation were the most significant variables influencing the distribution of invertebrate classes across the sampling range based on type I and III tests (Fig. 4.4). NH₄⁺, organic C, texture, titanite, soil temperature, chlorite, micas, P, gravel content, feldspar, and kaolin / chlorite also influenced the distribution of invertebrates to some extent (based on VIFs), apart from K. Correlation between invertebrate occurrences and conductivity (as a proxy for salt concentrations) therefore suggested that terrain age indirectly influences the distribution of biota. Detailed results regarding model selection, VIFs, final model evaluation, Type I, and Type III tests are provided in the supplemental materials. Class Chromadorea (Nematoda) occurred predominantly in areas with above-average S or conductivity (Fig. 4.4a, c). Monogonata (Rotifera) were not associated with a specific combination of variables (Fig. 4.3b). Enoplea (Nematoda) and arachnids (Chelicerata) occurred predominantly in low and flat areas with above-average levels of organic carbon, and phosphorus (Fig. 4.4). Heterotardigrades occurred frequently in areas with low salinity (Fig. 4.4a) and on sloped areas

rich in NH_4^+ (Fig. 4.4b, c). Bdelloid rotifers were associated with above-average elevation and high salinity (Fig. 4.4b).

4.4. Discussion

Our results indicate that soil salinity, and / or co-correlated variable S, are the most important constraints on Antarctic invertebrate biodiversity. PCA of environmental variables resulted in region-specific separation of Mount Menzies, Mawson Escarpment and Lake Terrasovoje, with cluster sizes seemingly corresponding to the spatial extent of sample collection (Fig. 4.1, Fig. 4.3a); such spatial separation was not observed for the biological data (Fig. 4.1, Fig. 4.3b). We found few variables significantly influenced invertebrate distribution patterns. The strong effect of electrical conductivity (and / or co-correlated variable S) was also pronounced in CCA (Fig. 4.4). Studies of the McMurdo Dry Valley soils suggested that salinity is an important factor influencing diversity of nematode communities (Freckman and Virginia, 1997; Powers et al., 1998; Treonis et al., 1999), with a similar effect also observed for mites (Elkins and Whitford, 1984). More complex communities were associated with younger weakly developed drifts with low conductivity in cold desert soils (Magalhaes et al., 2012). Cold desert soils (Mount Menzies, Mawson Escarpment) are dominated by the age-related accumulation of soluble salts from atmospheric deposition and weathering due to lack of precipitation (Bockheim, 1997; Campbell and Clardidge, 1987), while this effect is less pronounced in the polar desert soils of Lake Terrasovoje (Tedrow, 1966). The age-salinity correspondence is more pronounced for the soils of Mount Menzies and Mawson Escarpment, while this effect is reduced at Lake Terrasovoje where factors other than soil salinity predominantly influence soil biodiversity. The PCA of biological data (Fig. 4.3b; also supplemental. Fig. 4.12, Fig. 4.13) indicates Chromadorea (Nematodes) and Monogonata (Rotifera) are able to survive in the cold desert soils of Mount Menzies, while the heterogenic set of environmental factors in the Mawson Escarpment region offers suitable living conditions for all analysed invertebrate classes. Lower salinity soils at Lake Terrasovoje also support Bdelloidea and Heterotardigrada, in addition to all other taxa.

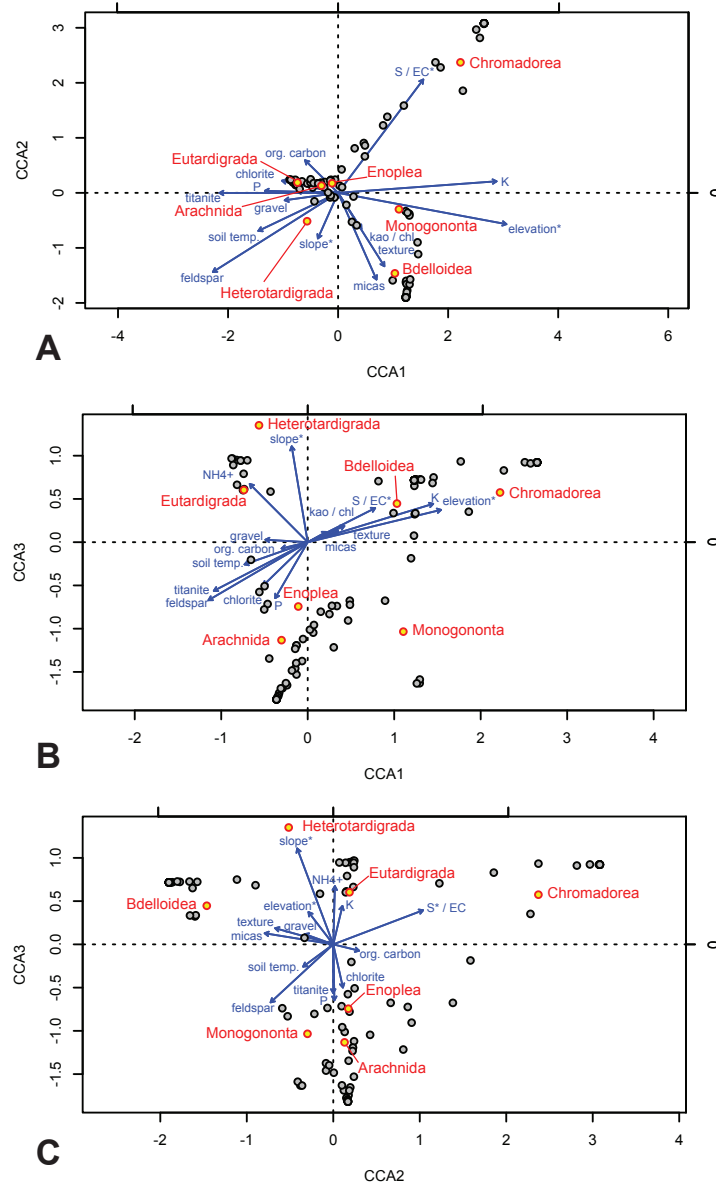


Figure 4.4.: Constrained Ordination using Canonical Correspondence Analysis of soil samples from the Prince Charles Mountains, East Antarctica. Shown are effects of 15 from 21 observation variables (in blue, co-correlated variables S and EC shown joined) contained in the highest-ranking model selected by Akaike Information Criterion ($AIC = 824.73$, $p = 0.021$). Observation variables and their vectors are related to seven invertebrate classes (red) found across sites (grey) found in proximity of each vector. Environmental vectors start from the mean value of the respective variable and end with the highest measurement for that variable (Palmer, 1993). Vectors are shortened if they reach out of the respective plane; hence all significant dimensions are plotted for comparison; with axes (A): CCA1 versus CCA2, (B): CCA1 versus CCA3, (C): CCA2 versus CCA3. Vectors marked with an asterisk (*) were significant in type I and/or type III tests after 1000 permutations (elevation, S, slope, and slope and S, respectively); variance inflation factors were below 10 for all variables, except K. Significances assigned to each axis were CCA1: $p = 0.001$, CCA2: $p = 0.001$, CCA3: $p = 0.001$ (1000 permutations). Biological and environmental observations were scaled symmetrically by square root of eigenvalues. Site labels were omitted, but no clear clustering into regions was observed. NH_4^+ omitted in (A) since vector is perpendicular to plane CCA1/CCA2.

Slope and elevation were not significant in both Type I and Type III CCA model tests and therefore regarded as proxies for other environmental variables. It has previously been suggested that spatial variables may constitute surrogates for relevant environmental variables at Cape Hallett (Sinclair et al., 2006). In line with this, co-variables linked to elevation have previously been difficult to determine for arctic tardigrades (Zawierucha et al., 2015). An elevational trend in the Antarctic was found to co-vary with soil properties such as carbon, nitrogen and salinity (Powers et al., 1998). In the Antarctic, elevation may be a strong proxy for salinity, as increasing elevation and lower temperature can increase the osmotic concentration of the soil solution, further inhibiting biological activity (Campbell and Clardidge, 1987; Freckman and Virginia, 1997; Powers et al., 1998). Slope has been suggested to have a localised effect on arctic soil biodiversity due to the increased moisture run-off (Zawierucha et al., 2015), however such an effect was deemed dependent on the moisture retention capabilities of local substrates (Sinclair et al., 2006). Hence, a relationship between slope angle and moisture should not be generalised without taking into account other variables, such as salinity. Also, spatial distribution of invertebrates is presumed only indirectly affected by the effect of slope or elevation. The distribution of invertebrate taxa in environmental space in the PCMs corresponds well with previous studies. Along an elevation gradient in Taylor Valley (McMurdo Dry Valleys) invertebrate biodiversity was greatest at the lowest elevation, where soil moisture, carbon, and nitrogen were highest, and salinity was lowest (Powers et al., 1998). In coastal locations of Victoria Land (Cape Hallett) variation in soil metazoan communities was related to differences in soil organic matter and moisture levels (Barrett et al., 2006). Salinity may affect soil invertebrate biodiversity by limiting the amount of available food or by directly affecting physiological functions (Nkem et al., 2006; Powers et al., 1998). At the same time, micro-eukaryotes may show a higher salinity tolerance than invertebrates since highly saline areas in the McMurdo dry valleys contained an abundant and complex mix of micro-eukaryotes (Fell et al., 2006) that may serve as food sources for Chromadorea and Monogonata in highly saline substrates. Our results and previous studies correspond well with the apparent split between Chromadorea and Monogonata, which are strongly associated with saline substrates (Fig. 4.3b) of Mount Menzies and Mawson Escarpment (Suppl. Fig. 4.12), and the arachnids, tardigrades and Enoplea, which are abundant in more nutrient-rich locations, predominantly at Lake Terrasovoje.

The observed split between nematode classes Enoplea and Chromadorea is also

concordant with previous studies. Nematodes were widely distributed and not correlated with moisture in the McMurdo Dry Valleys (Freckman and Virginia, 1997; Treonis et al., 1999), and there, nematode densities were also higher in dry soils (Powers et al., 1998). Environmental constraints may affect nematode species of the class Chromadorea differently (i.e. *Scottinema lindsayae* was deemed negatively correlated with soil moisture and C content; *Plectus antarcticus* dominantly found wet soils with low conductivity Barrett et al., 2006). At the same time, the abundance of *S. lindsayae* (Chromadorea) was negatively correlated with soil moisture and organic C content (Barrett et al., 2006; Freckman and Virginia, 1997) while *Eudorylaimus antarcticus* (Enoplea) was associated with high moisture and organic C (Freckman and Virginia, 1997). The dominance of Chromadorea in high-salinity areas may be due to species-specific physiological adaptations (Nkem et al., 2006), which allow species such as *S. lindsayae* to be almost ubiquitously distributed (Powers et al., 1998), if not affected by predation in moist and nutrient rich areas (Treonis et al., 1999). Enoplea (*Eudorylaimus*) may be less desiccation tolerant or dependent on a food source that is only available in moister soils (Powers et al., 1998).

In this study, tardigrades were associated with nutrient-rich soils along with arachnids and Enoplea, while rotifers were seemingly not strongly influenced by any environmental factor. Tardigrades were previously found to be associated with wet sediments (Barrett et al., 2006; Freckman and Virginia, 1997) and ornithogenic nutrient deposits (Zawierucha et al., 2015), and the majority of Antarctic mite species have similar habitat preferences (Pugh, 1993). Ornithogenic deposits are a major nutrient source of Antarctic soils, consequently nutrients are more abundant in coastal proximity (Bockheim, 1997). Tardigrades and rotifers were found associated with higher soil moistures (Freckman and Virginia, 1997), and found ubiquitously in coastal and inland sites (Barrett et al., 2006; Sohlenius and Bostroem, 2008; Sohlenius et al., 1996). While the literature regarding environmental requirements of mites, tardigrades and rotifers appears sparse, our results confirm the presence of these taxa in nutrient-rich substrates with low salinity (Fig. 4.3b, Fig. 4.4; also suppl. Fig. 4.12, Fig. 4.13). Antarctic rotifers are likely to be ubiquitously distributed across the sampling area and for this reason are not easily placed into the environmental space determined by the variables analysed here, apart from an apparent split also related to salinity.

Two of the most dominant environmental variables, temperature and water availability, have not been included in the analysis, due to practical constraints of retrieving

representative (i.e. time series) data from all study sites. Spot measurements of water are a poor proxy for actual water availability at a given sampling site, the same applies analogously for temperatures (Sansom, 1989; Sinclair et al., 2006). Inclusion of these environmental parameters will require detailed time series data for the considered areas in the Prince Charles Mountains, and a more extensive, future modelling approach, for example through species distribution modelling. Several possible approaches, with an Antarctic focus, are reviewed elsewhere (Chown et al., 2015; Gutt et al., 2012)

4.5. Conclusion

Linking biodiversity to environmental gradients is difficult in Antarctica, and CCA might be the most useful technique to investigate such relationships. It was previously found that various interrelated soil factors such as soil moisture, salinity, and pH may modify the effects of soil carbon and nitrogen on biodiversity, and that all variables collectively define suitable or inhospitable habitats for Antarctic soil invertebrates, both at local and more regional levels (Courtright et al., 2001; Freckman and Virginia, 1997; Powers et al., 1998). Yet, several studies could not find a clear link between environmental variables and invertebrate biodiversity or distribution, possibly because terrestrial Antarctica is characterized by limited ice-free ground and discontinuous substrates (Convey et al., 2014). The abiotic environment was shown to have some effect on the distribution of invertebrates on Ross Island, but vegetation and soil carbon did not explain the distribution of many taxa on Ross Island (Sinclair, 2001). Similarly, no correlations between animal or bacterial abundance, organic content and C/N ratio were found in soils of Basen Nunatak (Dronning Maud Land; Sohlenius et al., 1996). Among eukaryotes, distribution patterns seemed to reflect range expansion into deglaciated areas (Lawley et al., 2004). In summary, detecting relationships between environmental variables and species distributions in Antarctica is challenging, since organisms exhibit multiple scales of variability and their distribution is strongly influenced by spatial autocorrelation (Caruso et al., 2010). Canonical correspondence analysis (CCA) has been deemed particularly useful for the distinction between environmental variables representing real gradients or unimportant factors for species distribution (Palmer, 1993). Also, CCA performed well with skewed species distributions, quantitative noise in species abundance data, unusual sampling designs, and situations where not all of the factors determining

species composition were known (Palmer, 1993). The large distances between the three sampling regions (Fig. 4.1) mitigates spatial autocorrelation when comparing large-scale trends; application of CCA was likely most appropriate for this study. In further modelling approaches, time series data of water availability and annual temperature regimes will be able to provide further detail on relationships between the invertebrate biota of the Prince Charles Mountains and their abiotic environment. We used HTS-derived metagenetic biodiversity information to elucidate the relationships between soil properties and invertebrate biodiversity in the Prince Charles Mountains, Antarctica. Using data obtained from 103 soil samples collected across three broad sampling regions, we found invertebrate distributions were significantly influenced by electrical conductivity. Correlations between measured salt concentration and terrain age indicated, that terrain age may indirectly constrain Antarctic terrestrial biodiversity distributions, particularly in (inland) regions with cold desert soils. It has previously been suggested that geo-glaciological events and the presence of past glacial refuges may be more important than latitudinal variations in climatic and environmental conditions in determining the large-scale distributions of most Antarctic terrestrial fauna (Caruso et al., 2010). In addition, we find that the unique properties of Antarctic soils strongly constrain the distribution of Antarctic terrestrial taxa, in particular for classes Enoplea (Nematoda), Bdelloidea (Rotifera) and the phyla Tardigrada and Arachnida, which are restricted to nutrient rich, low-saline, (predominantly) more coastal sites. We demonstrated the value of HTS for exploring drivers of cryptic biodiversity at remote, largely unsurveyed sites across large spatial extents. Our approach is likewise applicable for other continental Antarctic taxa and regions.

4.6. Data accessibility

QIIME generated .biom files with raw invertebrate observations, reference trees and sequences, pre-processing shell scripts, environmental data, R analysis code and the corresponding R workspace file are available via <https://zenodo.org/record/19181>¹. Shell and R scripts are also provided in this document, Appendix C.

¹Data pre-release with closed access. Data will be publicly after manuscript publication.

4.7. Authors contributions

P.C. participated in field work; handled samples during extraction and sub-sampling; arranged soil geochemical analysis and logistics for mineral analysis; conducted laboratory work for mineral analysis; designed and applied molecular experiments and analysis approaches; interpreted results; designed and structured draft manuscripts; wrote manuscript; designed and created figures and tables. M.S. obtained funding for, and organised, planned, coordinated field campaign in Antarctica; contributed ideas to experiment and analysis design, interpretation of results, and manuscript structure; revised and edited draft and main manuscripts; contributed ideas for figures and tables; facilitated access to laboratory facilities, equipment and reagents for soil analysis. D.W. provided equipment for mineral analysis; processed raw mineral data for inclusion into downstream analyses; provided information and comments regarding geology of the sampling area; correlated terrain ages with environmental variables. L.C. contributed ideas on experiment and analysis design, interpretation of results, and manuscript structure; revised and edited draft and main manuscripts; contributed ideas for figures and tables. A.C. obtained funding, provided access to laboratory facilities, equipment and reagents for molecular analysis; provided and facilitated access to computational infrastructure for data analysis.

4.8. Acknowledgements

Antarctic samples collected and imported into Australia as regulated by DAFF permits ATEP 11-12-2355, IP12001186 and IP12001560. We thank Stephen Pederson, Greg Guerin and Jonathan Tuke (University of Adelaide) for helpful discussions on analysis methods and coding, Jimmy Breen (University of Adelaide) for maintaining the computational infrastructure required for sequence processing, and the members of the Australian Centre for Ancient DNA for helpful comments on the analyses. We also thank the members of the field party, Fiona Shanhun, Adrian Corvino, Josh Scarrow and Nick Morgan. We are grateful for the support provided by Helicopter Resources Pty. Ltd (TAS) for their invaluable support during the field campaign. We are indebted for the efforts of Perry Andersen, Michael Denton, and Bob Heath of Kenn Borek Air Ltd. in supporting our field campaign. We appreciate the help and support provided by the staff at Davis Station during the field campaign.

4.9. Funding

P.C. was supported by The University of Adelaide through an International Post-Graduate Research Scholarship. D. W. was supported by The University of Canberra. L.C. was supported through Australian Research Council linkage grant LP0991985. M.S. received funding from The Australian Antarctic Division, science project 2355. A.C. and M.S. received funding for this project through Australian Research Council linkage grant LP0991985.

4.10. Supplemental information: Salinity gradients determine invertebrate distribution in the Prince Charles Mountains, East Antarctica

4.10.1. Phylotype data generation for 18S and COI

Extractions were performed using a method optimised for the retrieval of DNA from different soil types and invertebrates Ophel-Keller et al. (2008); Pankhurst et al. (1996) that processes 400 g of starting material. Extraction, PCR and environmental controls were included in the molecular workflow as detailed below. Primer sequences for 18S sequencing on the Illumina MiSeq were sourced from (Gilbert et al., 2010; Parfrey et al., 2014). Primers “HCO2198” Folmer et al. (1994) and “mlCOIintF” Leray et al. (2013) were chosen for COI sequencing using the 454 GS FLX. Primer design and assignment is detailed below and in Fig. 4.5, Fig. 4.6. Illumina primer tags were sourced from Parfrey et al. (2014). Long extension times during amplification were used to counteract chimera formation Lenz and Becker (2008); Yu et al. (2012). All PCRs were carried out in triplicates to alleviate amplification biases (Gilbert et al., 2010). Amplicons above 0.25 ng/µl were pooled by weight for each marker. 18S libraries were paired-end sequenced in two runs on the MiSeq using 300 cycles. COI libraries were sequenced on two separate quarters of a 454 PicoTiterPlate. SILVA reference data (Pruesse et al., 2007) release 111 was used for taxonomy assignment to 18S phylotypes. COI reference data was compiled from earlier Antarctic studies (Velasco-Castrillón et al., 2014) and GenBank (Benson et al., 2011). Deconvolution, chimera screening, clustering, filtering of phylotypes, and naming and retention of Antarctic invertebrates was performed using QIIME version 1.8 (Caporaso et al., 2010) and R (R Development Core Team, 2011), including phylotypes obtained from control reactions and NCBI taxonomy nomenclature. All steps of phylotype data generation are further detailed elsewhere (Chapter 3). Design of molecular work and control reactions Antarctic soil DNA extracts from the Prince Charles Mountains were received in seven micro-titer plates, and then distributed between four plates, two with identical sample allocation for amplification using 18S or COI fusion primers. Several wells of each plate were reserved for control reactions. Controls included aliquots of blank extracts, H₂O (prior to amplification) or blank PCRs (after amplification) and several positive controls (Fig. 4.5). Methods for removal of contamination and retention of contamination are detailed elsewhere (Chapter

3). Phylotype data generated from Australian soil controls helped identification of Antarctic phylotypes in the final data: Only Antarctic phylotypes without linkage to Australian phylotypes were included into analyses for the current study. Phylotype information obtained from extraction controls or PCR controls was subtracted from Antarctic phylotypes using QIIME scripts, yielding a dataset deemed free of contamination (see also Appendix C).

4.10.2. Sequence tag selection and amplicon orientations

Biological samples processed in this study were part of a metagenetic sequencing workflow consisting of 192 samples (Fig. 4.6). While sample multiplexing was possible on the Illumina platforms with readily designed sequence tags Caporaso et al. (2012), only 140 adequate recognition sequences were available and tested for multiplexing on the 454 platform (Faircloth and Glenn, 2012; Roche, 2009a,b). Consequently, 454 fusion primers were specifically designed for the metagenetic workflow employed here and assigned to samples as listed in Fig. 4.6. Some combinations of adapter, recognition and primer sequences may inhibit amplification (Vallone and Butler, 2004), the selection of 454 fusion primers designed and employed here featured only sequences that did not exhibit hairpins (tested at <http://www.thermoscientificbio.com/webtools/multipleprimer/>). This further reduced amount of possible 454 fusion primers required sequencing of amplicons in two orientations to allow deconvolution, as shown in Fig. 4.6. Sequence constructs of the first sequencing orientation consisted of a primer with 454 adaptor “Lib-L Primer A key” (Roche, 2009b), recognition sequences unique to sample (Fig. 4.6) as well as the primer “mlCOIintF” (Leray et al., 2013); furthermore of a primer linking the reverse complemented 454 adaptor “Lib L Primer B key” (Roche, 2009b) with the reverse complemented sequence “HCO2198” (Folmer et al., 1994). Sequence constructs of reversed sequencing orientation contained the 454 adaptor “Lib-L Primer A key” (Roche, 2009b), recognition sequence unique to sample (Fig. 4.6) as well as primer “HCO2198” Folmer et al. (1994) in the first sequencing oligonucleotide. The second oligonucleotide contained the reverse complemented 454 adaptor “Lib L Primer B key” (Roche, 2009b), and “mlCOIintF” Leray et al. (2013).

4.10.3. Intermediate results of environmental data processing

Raw and pre-processed soil geochemical observations are summarised in Tab. 4.2 and Tab. 4.3. A first PCA showed to be impaired by outliers (Fig. 4.9), hence pre-processing was conducted. Removal of covariate variables is documented in Fig. 4.8. Geochemical and mineral variables for final PCA included gravel, texture, ammonium, nitrate, P, K, S, organic C, pH_{H₂O}, quartz, feldspar, titanite, pyroxene / amphibole / granite, micas, dolomite, kaolin / chlorite, calcite, and chlorite, elevation, slope and soil temperature.

4.10.4. Intermediate results of biological data processing

Raw, taxon corrected, phylotype data for each marker are plotted on the phylum level (Fig. 4.10). Both markers detected phyla Chelicerata, Nematoda and Rotifera; tardigrades were retrieved exclusively by 18S (Fig. 4.10). Phylotype tip agglomeration condensed phylotype numbers from 1860 to 29 for 18S, and from 37 to 31 for COI, respectively (Fig. 4.11). Combined invertebrate phylotype observations were defined for two super-phyla, 4 phyla, 7 classes, 16 orders, 19 families, 14 genera and 12 species (Tab. 4.4). Phylotype counts were agglomerated to class ranks Arachnida (Chelicerata), Bdelloidea and Monogonata (Rotifera), Chromadorea and Enoplea (Nematoda) and Eu- and Heterotardigrades (Tardigrada), and are shown in Fig. 4.12.

4.10.5. Intermediate results of biological data in relation to environment

Variables included in the final CCA model included elevation, sulphur, slope, ammonium, texture, titanite, soil temperature, organic carbon, chlorite, micas, potassium, phosphorus, gravel, feldspar, and kaolin / chlorite. For this highest-ranking model stepwise addition of variables to the unconstrained model towards the fully constrained model resulted in a drop of AIC values from 871 to 824 (Tab. 4.5) indicating improvement of predictive strength. VIFs ranged from min: 1.445 (ammonium), 1st Qu.: 1.912, median: 2.499, mean: 3.247, 3rd Qu.: 3.687 to max: 10.240 (potassium) and were below 10 for all but variable potassium (Tab. 4.6), indicating that all included variable except potassium contained independent information and little co-correlation. The final CCA model with the selected 15 variables was significant with $p = 0.021$. Significant information was loaded on 4 axes, with CCA1 – $p = 0.001$,

CCA2 – $p = 0.001$, CCA3 – $p = 0.001$ and CCA4 – $p = 0.028$. The first 3 axes were plotted in Fig. 4.4. Type I tests identified significant variables elevation ($p = 0.001$), sulphur ($p = 0.007$) and slope ($p = 0.082$). No re-arranged Type I tests were conducted, since the observed VIFs were below 10 for 9 of 10 variables. Type III tests identified significant variables S ($p = 0.007$) and slope ($p = 0.029$). Elevation was not significant in Type III tests due to its relatively high VIF (5.87).

4.10.6. Data and analysis scripts, additional figures and tables

Data and analysis scripts can be accessed at <https://zenodo.org/record/19181>. Analysis source code is provided in this document (Appendix C).

	1	2	3	4	5	6	7	8	9	10	11	12
A	AC24981 Reinbolt Hills	AC23007	AC23008	AC23009	AC23010	AC23011	AC08310 Mount Menades	AC26651	AC08312	AC08313	AC08314	PCR (-)
B	AC26632	AC08315	AC08316	AC08317	AC26623	AC08319	AC26619	AC08320	AC08323	AC08324	AC26616	PCR (-)
C	AC08325	AC24977	AC08326	AC08327	AC26633	AC08328	AC23016		AC26625	AC08332	AC26645	PCR (-)
D	AC08340	AC29096	AC29038	AC08341	AC29039	AC29093	AC08343	AC26644	AC29097	AC29049	AC29043	PCR (-)
E	AC29045	AC08347	AC29046	AC29048	AC23022	AC29099	AC23023	AC08351	AC29102	AC29050	AC24980	PCR (-)
F	AC29051	AC08353	AC29103	AC26626	AC23025	AC23024	AC08355	AC29104	AC08356	AC29052	AC29105	PCR (-)
G	AC29106	AC08358	AC29053	AC29107	AC29130	AC29054	AC08361	AC29108	AC08362	AC29132	AC08363	PCR (-)
H	AC29055	AC08364	AC29056	AC08365	AC29110	AC08366	AC24982	AC23026	AC08367	AC08401 soil ctrl 1	AC29126 soil ctrl 2	PCR (-)

	1	2	3	4	5	6	7	8	9	10	11	12
A	AC29111	AC29112	AC08368	AC29113	AC29131	AC23027	AC29128	AC29057	AC23030	AC24957	AC29114	PCR (NTC)
B	AC23028	AC29115	AC08441	AC29116	AC23035	AC23031	AC08464	AC08371 Amery Oasis	AC24959	AC24959	AC08372	PCR (NTC)
C	AC08373	AC24958	AC29039	AC26621	AC08375	AC26620	AC08376	AC23038	AC08377	AC26629	AC26649	PCR (NTC)
D	AC08378	AC24961	AC08379	AC26648	AC24962	AC08380	AC24978	AC08381	AC26647	AC08382	AC26617	PCR (-)
E	AC08383	AC24963	AC26636	AC26624	AC08494	AC08385	AC24965	AC08384	AC08386	AC26640	AC26618	PCR (-)
F	AC08388	AC08389	AC24970	AC08390	AC24968	AC24969	AC29061	AC29062	AC08394	AC29063	AC08395	PCR (-)
G	AC29064	AC08398	AC29069	AC29070	AC29072	AC29074	AC29075	AC24972	AC24975 in via	AC29077	AC29118	PCR (-)
H	AC29123	AC29041	AC08413 R 3920-2 E11	AC08427 T 3920-2 A3	AC08415 N 3920-2 E6	AC08407 M 3920-03	H3 - H6 mixed 28.12.11 ins ctrl 1 (+)	insect blend ins ctrl 1 (+)	damselfly ins ctrl 2 (+)	AC08401 soil ctrl 1 (+)	AC29126 soil ctrl 2 (+)	PCR (-)

Figure 4.5.: DNA extracts allocation on micro-titer plates. For each marker (18S/COI) an identical layout was chosen. Sample origins were indicated with shades of yellow, given were extract identifiers (“AC”). Sample names used in this study corresponded to plate positions (ranging from 1.1.A to 2.12.H.; for plate 1, well 1A, to plate 2, well 12H, respectively). Row 12 of each plate contained pools of extract controls or PCR controls (H₂O prior to amplification). Wells 10H and 11H on plate 1 and wells 3H – 11H of plate 2 contained control DNA extracts, of which wells 8H, 10H and 11H were analysed elsewhere (sec. 3.8). Additionally, phylotypes obtained from wells 10H and 11H were used for the retention of Antarctic invertebrates as detailed elsewhere (sec. 3.8).

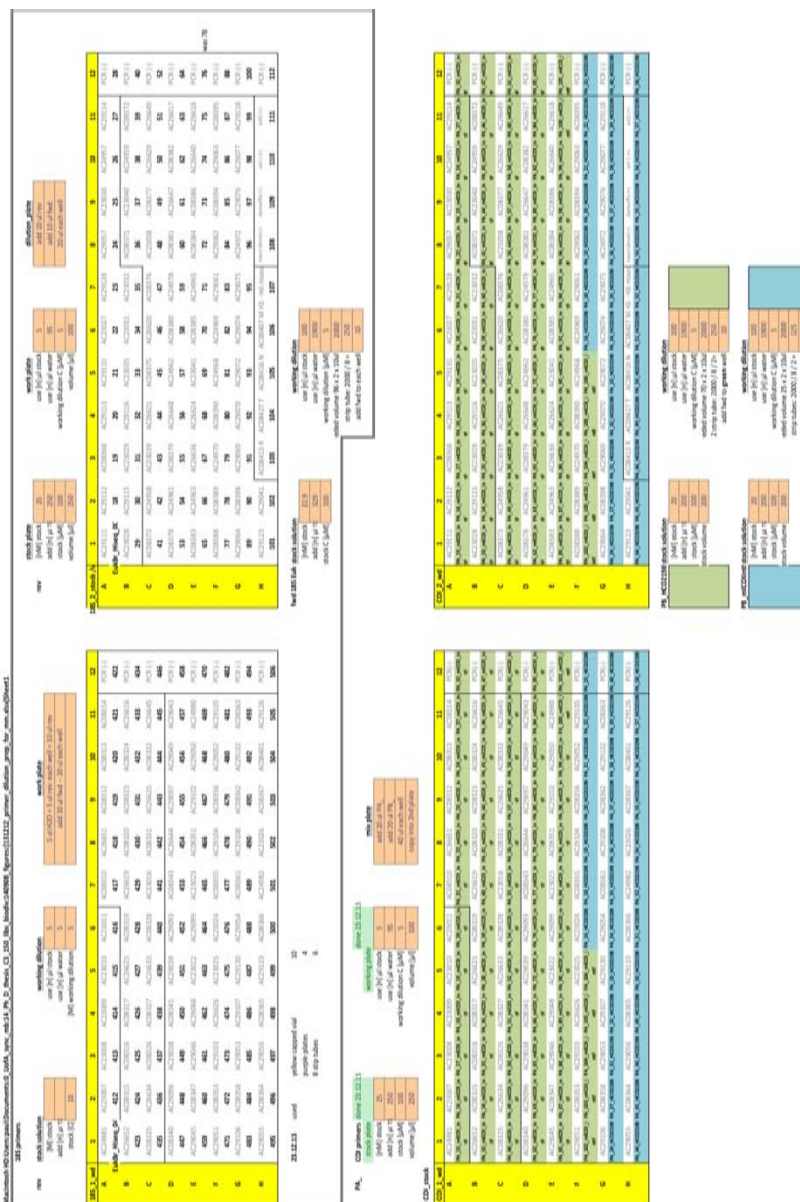


Figure 4.6.: Fusion primers employed for sequencing of 18S amplicons on the Illumina platform, and for COI amplicon sequencing on the 454 platform. Illumina fusion primers (upper) employed in this study were described in detail elsewhere (chapter 2), numbers over wells refer to barcode identifiers in (Parfrey et al., 2014). 454 fusion primers (lower) are named by their forward adapter, “MID” number (Roche, 2009a,b) and the employed forward primer. Colours blue and green reflect amplicon orientation during sequencing. Also provided are the respective 454 reverse primers, named after the same convention.

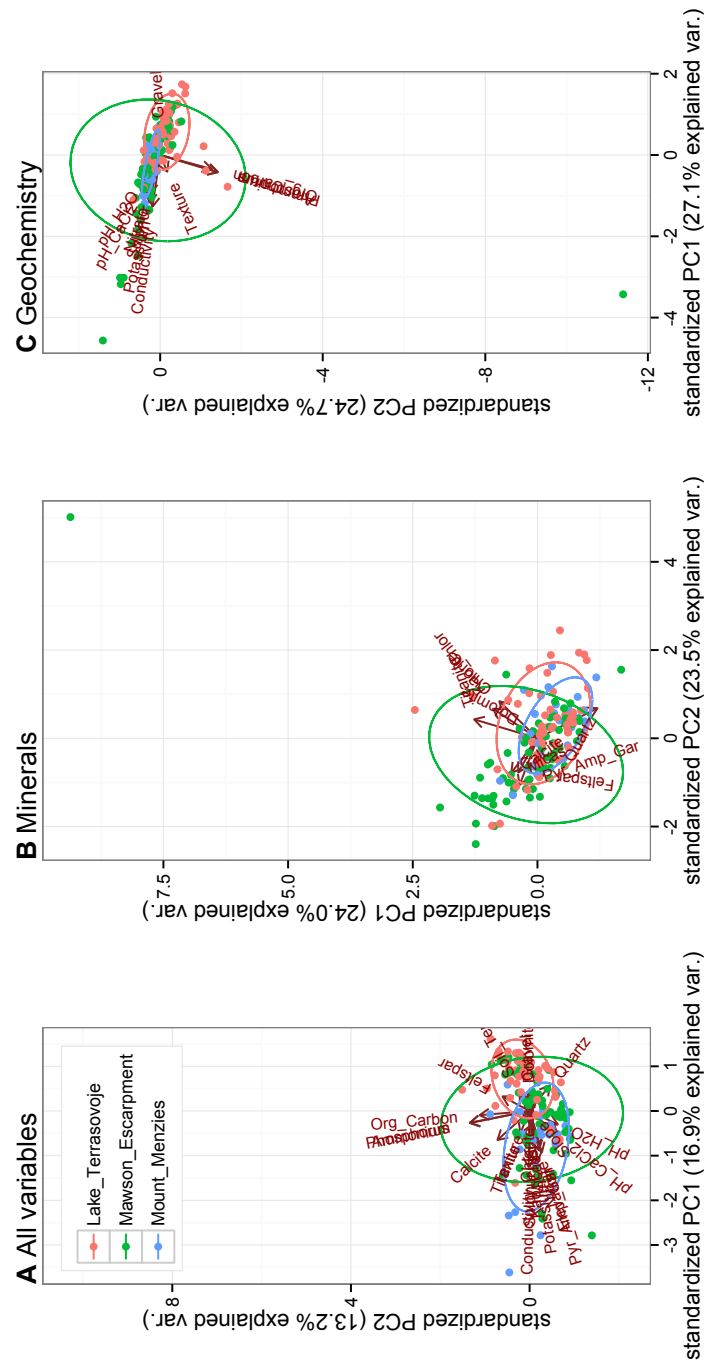


Figure 4.7.: Initial Principal Component Analysis (PCA) of mineral and geochemical observations. Initial PCA of combined observations data revealed necessity for pre-processing. PCA was calculated with scaled and centred mineral and geochemical observations available for 141 sites, sampling regions were coded with colours. (A): Outliers in combined data required further processing. (B) and (C): both mineral and geochemical data required removal of outlying data and pre-processing.

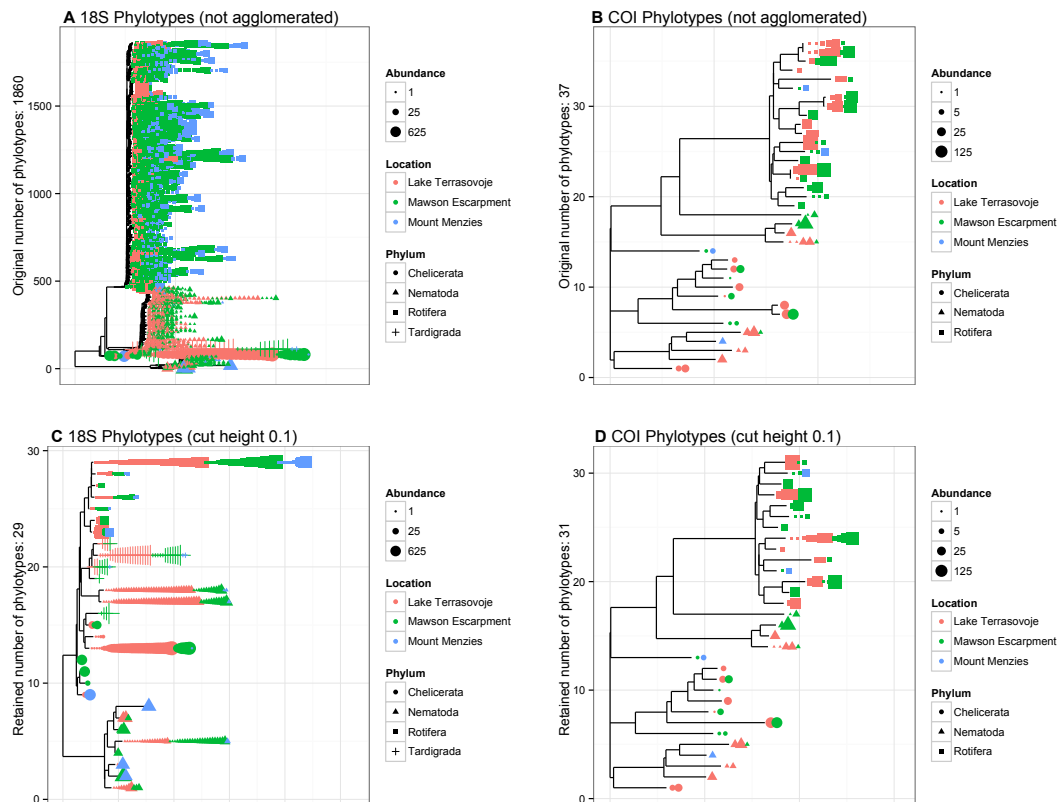


Figure 4.11.: Tree tip agglomeration. Tree agglomeration was chosen a taxon-agnostic means to decrease site heterogeneity among biological data. Invertebrate reference trees were calculated from all Antarctic invertebrates phylotypes contained in 18S and COI data (A and B, respectively). Each phylotype is referenced to its sample of origin (colours) and a taxonomic assignment (shapes, here shown on phylum level). Agglomeration of tree tips for 18S (C) and COI (D) reduced the number of phylotypes while site references were maintained, practically decreasing the site heterogeneity among biologic observations. 18S data retrieved information for four Antarctic invertebrate phyla (C), for COI information for three phyla could be obtained (D). Basal branches remained unresolved for reasons of computational efficiency, without affecting the results of this data processing step.

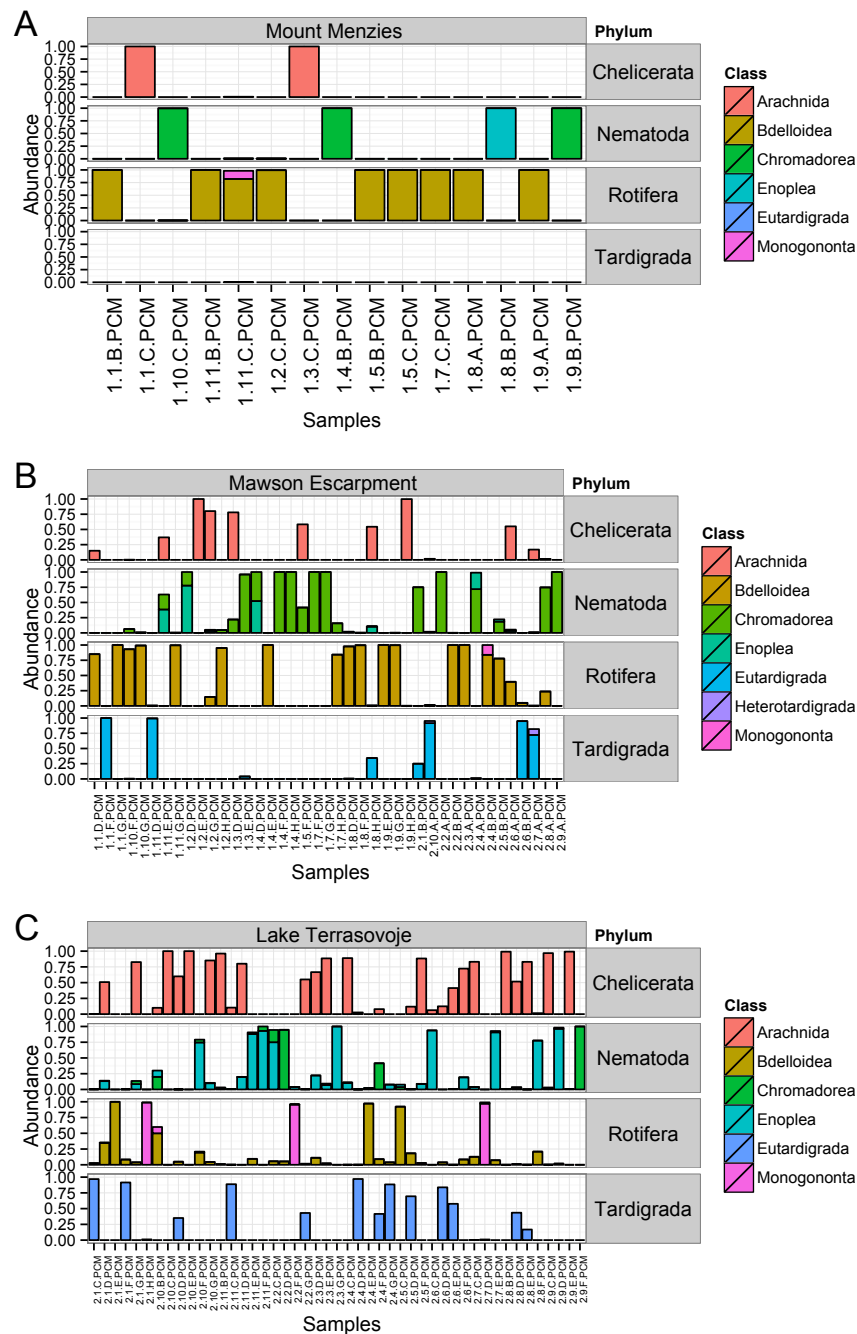


Figure 4.12.: Phylum and class level assignments of invertebrate phylotypes shown by sampling region. Phylotype information was combined from 18S and COI data. Taxonomic information for all phylotypes is available on the phylum and class level (facets and colours, respectively). Location references of phylotype data are here used to group data by sampling region Mount Menzies, Mawson Escarpment and Lake Terrasovoje (A, B and C, respectively). (A) At Mount Menzies nematodes and rotifers constitute the majority of biodiversity, mites occur in places, and tardigrades were not detected. (B) In the Mawson Escarpment, mites and tardigrades are more abundant. (C) At Lake Terrasovoje, mites and tardigrades are highly abundant, and all phyla exhibited the highest diversity (on class level).

Table 4.1.: Extraction methods employed for soil geochemical analysis, provided by CSBP Soil and Plant Analysis Laboratory (Bibra Lake, AU-QLD) and referenced in Rayment and Lyons, 2011.

Analytes	Method, Unit, Detection Limit
P and K	Method 9B1; Unit: mg/kg; Limits: 2
Soil pH in CaCl ₂	Method 4B1; Unit: pH; Limits: 2 decimal points
Soil pH in H ₂ O	Method 4A1; Unit: pH; Limits: 2 decimal points
Organic C	Method 6A1; Unit: %; Limits: 0.05
NH ₄ ⁺ and NO ₃ ⁻ N	Method 7C2; Unit: mg/kg; Limits: 1
KCl S	Method 10D1; Unit: mg/kg; Limits: 0.5
electric conductivity	Method 3A1; Unit: dS/m; Limits: 0.01
P	Method 9C2; Unit: mg/kg; Limits: 0.1

Table 4.2.: Summary of soil geochemical and mineral measurements before pre-processing.

	Gravel (%)	Texture	NH ₄ ⁺ (mg/kg)	NO ₃ ⁻ (mg/kg)	P (mg/kg)	K (mg/kg)	S (mg/kg)	C (%)
Min.	0	1	0	0	0	28	0.7	0
1st Qu.	5	1	0	0	0	110	5.7	0
Median	15	1	0	3	2	261	28.5	0.1
Mean	20.5	1.411	6.773	106.4	37.6	322	690.7	0.176
3rd Qu.	30	2	2	56	6	449	367.4	0.14
Max.	75	4	803	2343	4580	1300	10695.4	4.5
	Conductivity (ds/m)	pH (CaCl ₂)	pH (H ₂ O)	Quartz	Feldspar	Titanite	Pyr. /Amp. /Gar.	Micas
Min.	0.01	4.4	4.7	0	0	0	0	0
1st Qu.	0.027	6.5	7	0.5308	0.1813	0.02781	0.007209	0.00756
Median	0.091	7.2	7.7	0.5977	0.2474	0.04721	0.033483	0.01864
Mean	0.6097	7.15	7.674	0.5978	0.2324	0.05194	0.041951	0.02951
3rd Qu.	0.787	7.8	8.4	0.6479	0.277	0.05747	0.064683	0.03669
Max.	6.583	9	9.9	0.8854	0.4228	0.79667	0.180997	0.17066
	Dolomite	Kao. / Chl.	Calcite	Chlorite	Elevation (m.a.s.l)	Slope (°)	Soil Temp. (°C)	
Min.	0	0	0	0.00E+00	11	0	-15.3	
1st Qu.	0.006198	0.001492	0	0.00E+00	195	1	1.2	
Median	0.016019	0.008958	0.004773	4.36E-06	817	4	7.2	
Mean	0.02156	0.013338	0.006853	4.67E-03	849.9	6.319	5.335	
3rd Qu.	0.02471	0.015338	0.010191	6.31E-03	1256	8	11	
Max.	0.389919	0.144297	0.044091	5.88E-02	2180	30	27	

Table 4.3.: Summary of soil geochemical and mineral measurements after pre-processing. Pre-processing was conducted through Box-Cox transformation, removal of outliers and removal of co-correlated variables. Co-correlated variables Conductivity and pH CaCl₂ removed; bold variables were used in Canonical Correspondence Analysis (CCA).

	Gravel	Texture	NH₄⁺	NO₃⁻	P	K	S	C
Min.	-1.0973	-0.7707	-0.5155	-0.3417	-0.5358	-2.05388	-1.701	-0.45551
1st Qu.	-0.8132	-0.7707	-0.5155	-0.3417	-0.5358	-0.85708	-0.8568	-0.45551
Median	-0.245	-0.7707	-0.5155	-0.3303	-0.3277	0.08594	-0.2089	-0.14152
Mean	0	0	0	0	0	0	0	0
3rd Qu.	0.3232	1.2719	0.4341	-0.13	0	0.74615	0.8203	-0.01592
Max.	2.8802	1.6501	5.6571	7.8226	6.6424	2.27722	2.0825	8.99551
	Conductivity	pH (CaCl₂)	pH (H₂O)	Quartz	Feldspar	Titanite	Pyr. /Amp. /Gar.	Micas
Min.	-	-	-2.27665	-2.45616	-2.4392	-1.941543	-1.083	-0.8834
1st Qu.	-	-	-0.74704	-0.59985	-0.7702	-0.783105	-0.8924	-0.649
Median	-	-	0.02919	0.04251	0.1366	0.003487	-0.1976	-0.3057
Mean	-	-	0	0	0	0	0	0
3rd Qu.	-	-	0.77766	0.4998	0.5992	0.44259	0.6227	0.2458
Max.	-	-	2.30212	2.40992	3.149	5.248783	2.6703	4.2086
	Dolomite	Kao. / Chl.	Calcite	Chlorite	Elevation	Slope	Soil Temp.	
Min.	-0.61196	-0.7774	-0.8838	-0.5685	-1.8947	-0.8896	-2.6439	
1st Qu.	-0.41159	-0.6839	-0.8838	-0.5685	-1.0185	-0.7363	-0.5139	
Median	-0.09408	-0.2159	-0.2434	-0.568	0.1846	-0.2765	0.2607	
Mean	0	0	0	0	0	0	0	
3rd Qu.	0.17463	0.183	0.4821	0.211	0.7372	0.3365	0.7384	
Max.	10.69372	6.2215	4.3085	6.3888	1.6703	3.4017	1.784	

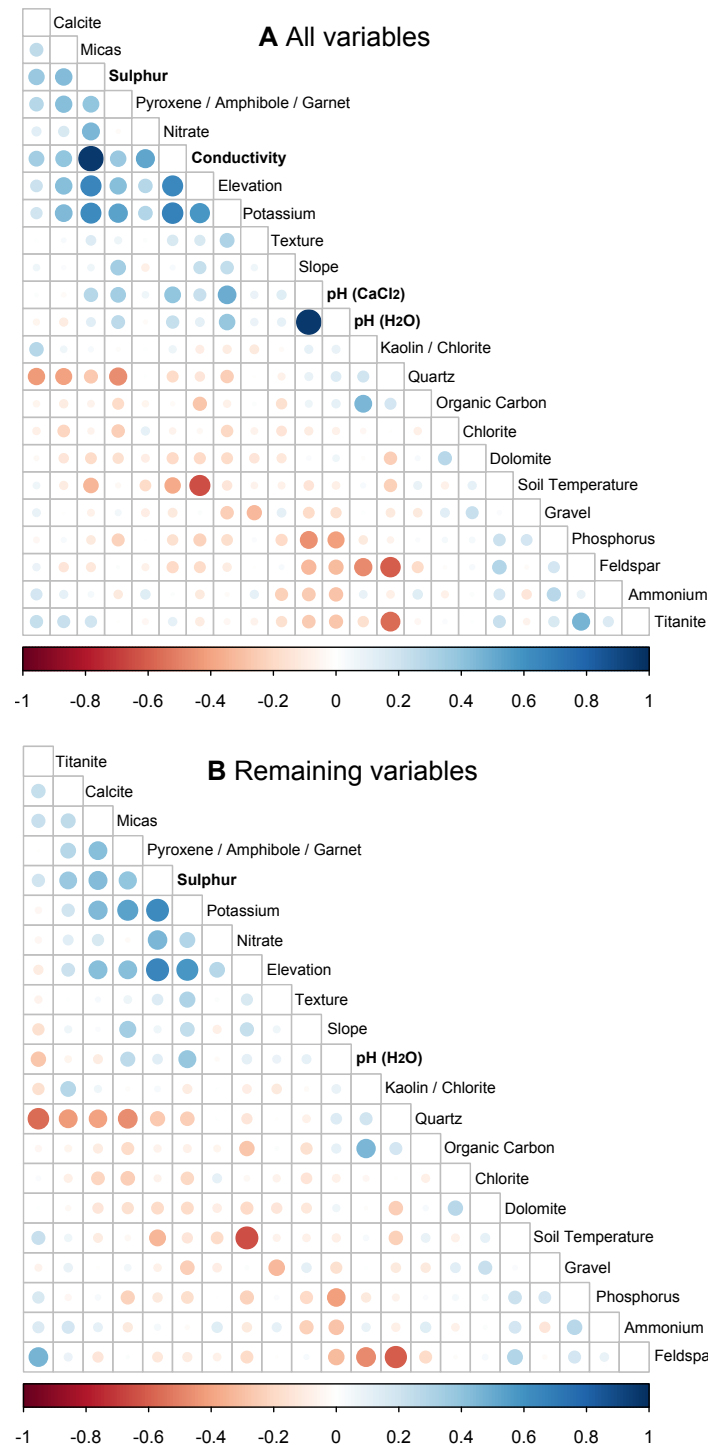


Figure 4.8.: Removal of co-correlated observations variables. (A): Among 23 observation variables initially contained in our data set, pairwise Pearson correlation (after scaling, centring and removal of outliers) was above 0.7 for variables ‘S’ and ‘Conductivity’, and pH of H₂O and CaCl₂, respectively. (B): Variables ‘Conductivity’ and ‘pH CaCl₂’ were removed from the dataset, and 21 variables remained.

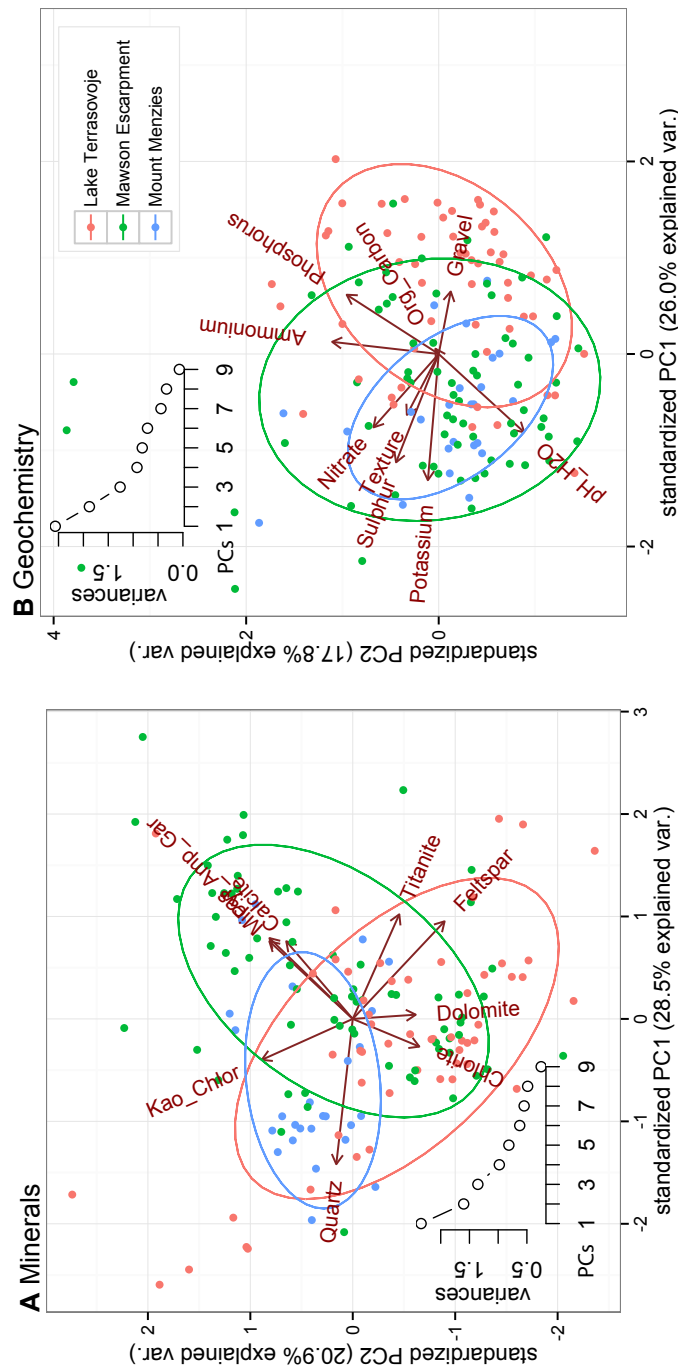


Figure 4.9.: Principal Component Analyses (PCA) of mineral and geochemical observations. (A): Mineral data and (B): Geochemical observations. PCA was calculated after pre-processing (i.e. Box-Cox transformation, centring, variable scaling to unit variance, removal of co-correlated variables and outliers), for observations across 141 sites. Sampling regions were coded with colours. Insets show variances of first 10 Principal Components (PCs). (A): In 9 principal components 7 contained 95% of all variance. (B): In 9 principal components 8 contained 95% of all variance.

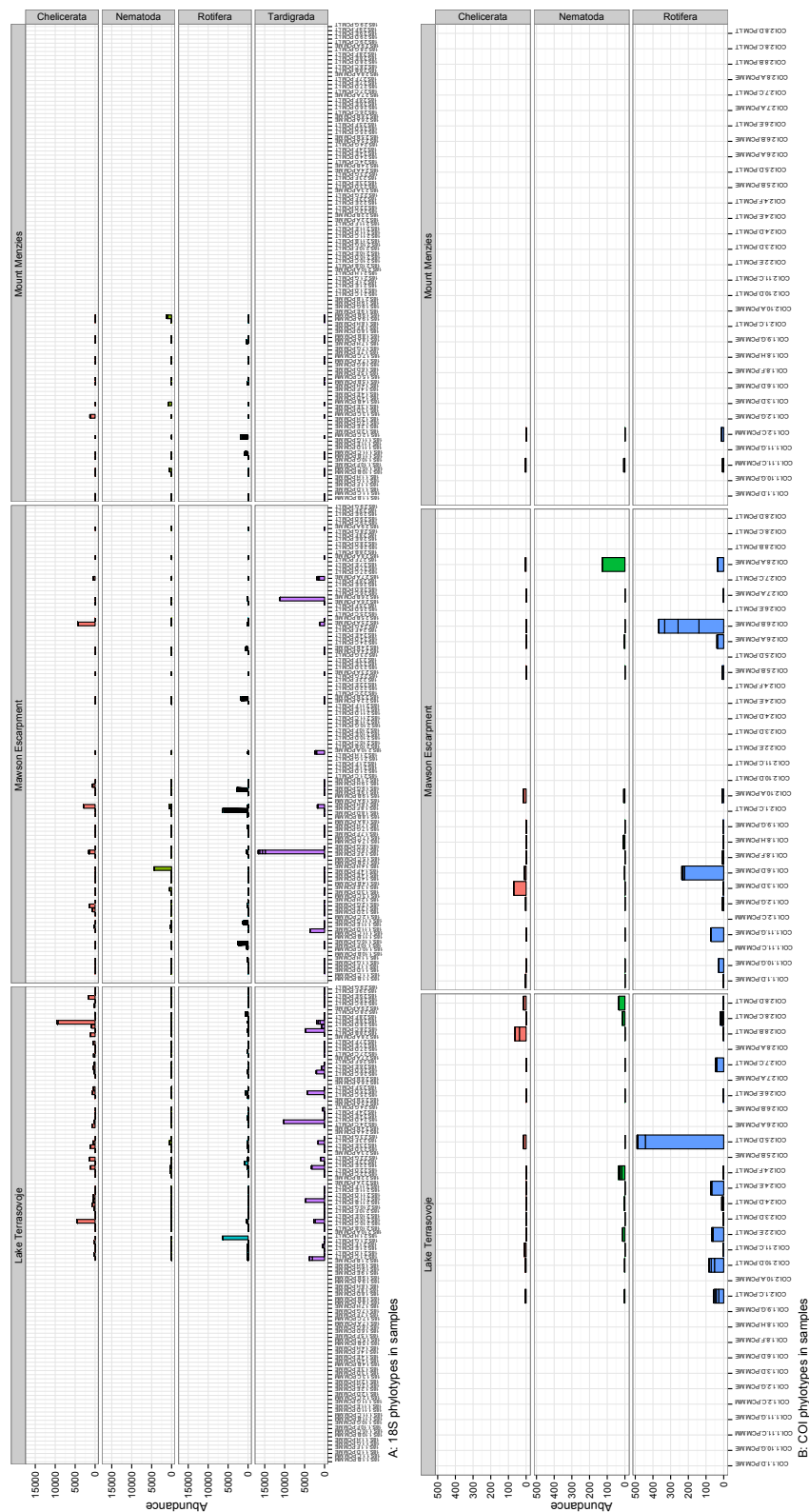


Figure 4.10.: Antarctic phylotype data obtained from both markers by phylum and sampling region. Shown are raw abundance values before tip and class level agglomeration.

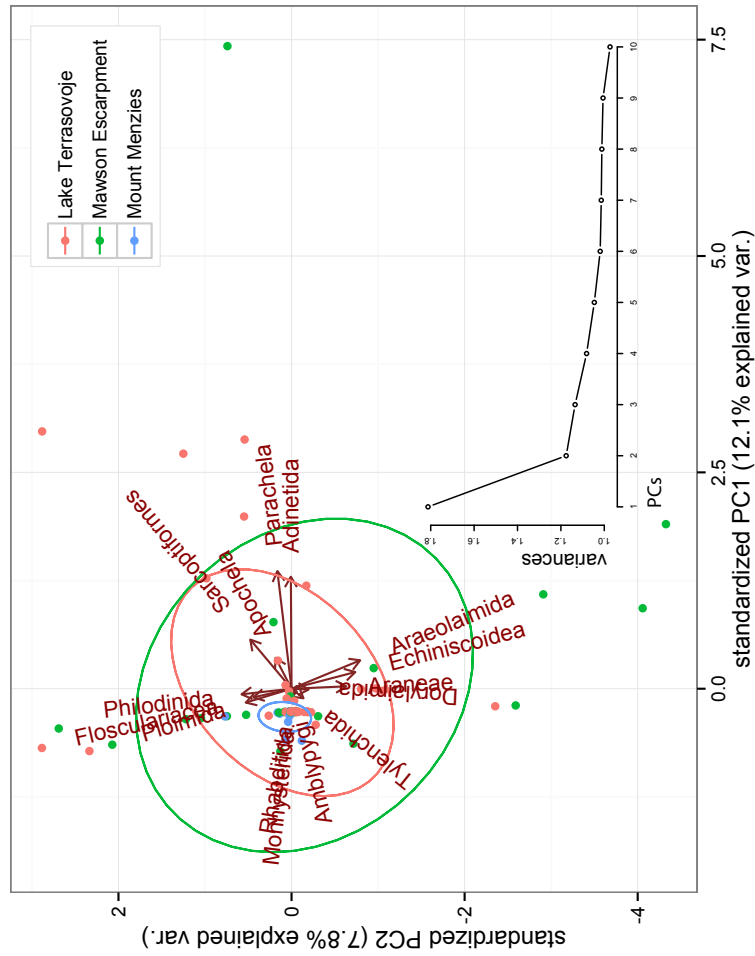


Figure 4.13.: Principal Component Analysis (PCA) of combined phylotype data. Shown are agglomerated order-level phylotypes. PCA was calculated after pre-processing (i.e. Box-Cox transformation, centring, variable scaling to unit variance) of combined invertebrate phylotype counts. Sampling regions were coded with colours. Inset shows variances of first 10 Principal Components (PCs). 14 of 15 PCs were needed to account for 95% variance.

Table 4.4.: Taxonomic composition and resolution of combined metagenetic data. Taxonomic composition and resolution of combined 18S and COI metagenetic data. Data combined after tree agglomeration for each marker; hence each phylotype identifier represents several highly similar phylotypes (*). For Canonical Correspondence Analysis (CAA) data was further agglomerated to the class level (**). Restriction to Antarctic phylotypes ensures that unlikely taxonomic assignments are caused by deficiencies in reference data, and do not represent contaminants (***) . Empty fields represent missing taxonomic information in the reference data (****).

Phylotype identifier (*)	Superphylum	Phylum	Class (**)	Order	Family	Genus	Species
ets_denovo43126	Ecdysozoa	Chelicerata	Arachnida	Amblypygi (***)	Phrynichidae	<i>Damon</i>	(****)
coi_denovo156	Ecdysozoa	Chelicerata	Arachnida	Araneae (***)	Linyphiidae	<i>Erigone</i>	<i>Erigone tiraldensis</i>
coi_denovo884	Ecdysozoa	Chelicerata	Arachnida	Araneae	Linyphiidae	<i>Meioneta</i>	<i>Meioneta simplex</i>
coi_denovo927	Ecdysozoa	Chelicerata	Arachnida	Araneae	Lycosidae	<i>Pardosa</i>	<i>Pardosa lapponica</i>
coi_denovo53	Ecdysozoa	Chelicerata	Arachnida	Araneae	Lycosidae	<i>Pardosa</i>	<i>Pardosa lapponica</i>
coi_denovo939	Ecdysozoa	Chelicerata	Arachnida	Araneae			
coi_denovo542	Ecdysozoa	Chelicerata	Arachnida	Araneae			
coi_denovo1377	Ecdysozoa	Chelicerata	Arachnida	Araneae			
coi_denovo240	Ecdysozoa	Chelicerata	Arachnida	Sarcoptiformes			
ets_denovo95019	Ecdysozoa	Chelicerata	Arachnida				
ets_denovo230284	Ecdysozoa	Chelicerata	Arachnida				
ets_denovo9425	Ecdysozoa	Chelicerata	Arachnida				
ets_denovo177020	Ecdysozoa	Chelicerata	Arachnida				
ets_denovo116385	Ecdysozoa	Chelicerata	Arachnida				
ets_denovo79840	Ecdysozoa	Chelicerata	Arachnida				
coi_denovo2	Ecdysozoa	Chelicerata	Arachnida				

Phylotype identifier (*)	Superphylum	Phylum	Class (**)	Order	Family	Genus	Species
coi_denovo475	Ecdysozoa	Nematoda	Chromadorea	Araeolaimida	Plectidae	<i>Plectus</i>	<i>Plectus cf. frigophilus</i>
coi_denovo496	Ecdysozoa	Nematoda	Chromadorea	Araeolaimida	Plectidae	<i>Plectus</i>	<i>Plectus cf. frigophilus</i>
coi_denovo137	Ecdysozoa	Nematoda	Chromadorea	Araeolaimida	Plectidae	<i>Plectus</i>	<i>Plectus murrayi</i>
ets_denovo185546	Ecdysozoa	Nematoda	Chromadorea	Araeolaimida	Plectidae		
ets_denovo209579	Ecdysozoa	Nematoda	Chromadorea	Monhysterida	Monhysteridae		
ets_denovo68302	Ecdysozoa	Nematoda	Chromadorea	Rhabditida	Cephalobidae		
ets_denovo176366	Ecdysozoa	Nematoda	Chromadorea	Rhabditida	Cephalobidae		
ets_denovo125623	Ecdysozoa	Nematoda	Chromadorea	Rhabditida	Panagrolaimidae	<i>Halicephalobus</i>	<i>Halicephalobus cf. gingivalis</i>
ets_denovo100206	Ecdysozoa	Nematoda	Chromadorea	Rhabditida	Rhabditidae	<i>Rhabditis</i>	
coi_denovo139	Ecdysozoa	Nematoda	Chromadorea	Tylenchida	Aphelenchoididae	<i>Bursaphelenchus</i>	<i>Bursaphelenchus cocophilus</i>
coi_denovo814	Ecdysozoa	Nematoda	Chromadorea	Tylenchida	Aphelenchoididae	<i>Bursaphelenchus</i>	<i>Bursaphelenchus cocophilus</i>
coi_denovo617	Ecdysozoa	Nematoda	Chromadorea	Tylenchida	Aphelenchoididae	<i>Bursaphelenchus</i>	<i>Bursaphelenchus cocophilus</i>
ets_denovo169860	Ecdysozoa	Nematoda	Chromadorea	Tylenchida	Dolichodoridae		
ets_denovo161289	Ecdysozoa	Nematoda	Chromadorea	Tylenchida	Hoplolaimidae	<i>Pratylenchus</i>	<i>Pratylenchus thornei</i>
coi_denovo536	Ecdysozoa	Nematoda	Enoplea	Dorylaimida	Longidoridae	<i>Xiphinema</i>	
ets_denovo75692	Ecdysozoa	Nematoda	Enoplea	Dorylaimida			
ets_denovo111717	Ecdysozoa	Nematoda	Enoplea	Dorylaimida			
ets_denovo10491	Ecdysozoa	Tardigrada	Eutardigrada	Apocheila	Milnesiidae	<i>Milnesium</i>	<i>Milnesium tardigradum</i>
ets_denovo40422	Ecdysozoa	Tardigrada	Eutardigrada	Paracheila	Macrobiotidae	<i>Macrobiotus</i>	<i>Macrobiotus hufelandi</i>
ets_denovo45114	Ecdysozoa	Tardigrada	Eutardigrada	Paracheila	Macrobiotidae		
ets_denovo224794	Ecdysozoa	Tardigrada	Eutardigrada	Paracheila	Macrobiotidae		
ets_denovo134265	Ecdysozoa	Tardigrada	Heterotardigrada	Echiniscoidea	Echiniscidae		

Phylotype identifier (*)	Superphylum	Phylum	Class (**)	Order	Family	Genus	Species
coi_denovo1236	Lophotrochozoa	Rotifera	Bdelloidea	Adinetida	Adinetidae	<i>Adineta</i>	<i>Adineta vaga</i>
coi_denovo1430	Lophotrochozoa	Rotifera	Bdelloidea	Adinetida	Adinetidae	<i>Adineta</i>	<i>Adineta vaga</i>
ets_denovo171283	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
ets_denovo97814	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
ets_denovo189992	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
ets_denovo4100	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
ets_denovo111523	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
coi_denovo699	Lophotrochozoa	Rotifera	Bdelloidea	Philodinida	Philodinidae		
coi_denovo773	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo583	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo200	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1230	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1362	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1409	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1623	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1145	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo217	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo1052	Lophotrochozoa	Rotifera	Bdelloidea				
coi_denovo662	Lophotrochozoa	Rotifera	Bdelloidea				
ets_denovo73075	Lophotrochozoa	Rotifera	Monogononta	Flosculariacea	Flosculariidae		
ets_denovo220647	Lophotrochozoa	Rotifera	Monogononta	Ploimida			

Table 4.5.: Summary of model selection process for the final model used in Canonical Correspondence analysis (CCA). In each step (Step) a variable was added to an unconstrained model towards a fully constrained model, and the result was tested with 999 permutations. Addition of a variable removed one degree of freedom (Df). For each step the residual degrees of freedom and Deviance (Resid. Df and Resid. Dev., respectively) were assessed. For the selected model, addition or removal of variables resulted in a final Akaike Information criterion of 824.73. Lower ranking models not shown.

Step	Df	Deviance	Resid. Df	Resid. Dev.	AIC
1	-	-	102	475580.8	871.07
2 + elevation	-1	56865.87	101	418715	859.95
3 + S	-1	44454.43	100	374260.5	850.39
4 + quartz	-1	28653.09	99	345607.4	844.19
5 + slope	-1	23439.02	98	322168.4	838.95
6 + NH_4^+	-1	16679.44	97	305489	835.48
7 + texture	-1	14470.92	96	291018	832.48
8 + titanite	-1	9675.71	95	281342.3	831.00
9 + pH (H_2O)	-1	6582.63	94	274759.7	830.56
10 + soil temperature	-1	6775.03	93	267984.7	829.99
11 + org C	-1	6450.50	92	261534.2	829.48
12 + Chlorite	-1	6498.16	91	255036	828.89
13 + Micas	-1	5538.27	90	249497.8	828.63
14 + K	-1	5843.47	89	243654.3	828.18
15 + P	-1	5521.46	88	238132.8	827.82
16 + gravel	-1	6996.76	87	231136.1	826.75
17 - pH pH (H_2O)	1	4415.63	88	235551.7	826.70
18 + Feldspar	-1	4913.89	87	230637.8	826.53
19 - Quartz	1	1212.16	88	231850	825.07
20 + Kao. / Chl.	-1	5196.19	87	226653.8	824.73

Table 4.6.: Variance inflation factors (VIF) for model employed in Canonical Correspondence Analysis (CCA). A value above 10 indicates that the corresponding variable does not contain independent information.

Variable	VIF
Elevation	5.87
Slope	2.23
S	2.69
NH ₄ ⁺	1.44
Texture	1.49
Titanite	2.92
Soil temp.	3.48
Org. C	1.48
Chlorite	1.81
Micas	2.01
K	10.24
P	2.5
Gravel	4.2
Feldspar	3.9
Kaol. / Chl.	2.44

5. Synthesis

Paul Czechowski¹

¹ Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005 Australia

5.1. Summary

5.1.1. Technical and computational methods

This work applied a wide range of technical and computational methods to increase the biological information available for remote ice-free regions of continental Antarctica and applied these to explore environmental constraints on Antarctic biodiversity. Initially, the possibility of using *metagenetic high throughput sequencing* (MHTS) to generate baseline biodiversity data was evaluated (Chapter 1). Subsequently, the eukaryotic diversity of three hitherto unsurveyed regions in the Prince Charles Mountains was explored (Chapter 2). In order to apply MHTS for the study of Antarctic invertebrates, taxonomic assignment fidelities between metagenetic markers and morphological approaches were compared (Chapter 3). Lastly, MHTS-generated biodiversity data was used to explore the environmental determinants of invertebrate distribution in the Prince Charles Mountains (Chapter 4).

5.1.2. Biodiversity information from the Prince Charles Mountains

Chapter 1 The first chapter of this thesis showed that current MHTS approaches mainly employ workflows coined *amplicon sequencing* that require relatively few laboratory steps, and for this reason are preferable to retrieve large-scale biodiversity information from terrestrial Antarctica. Such approaches have previously been

applied in Antarctica (e.g. Bottos et al. 2014; Makhalanyane et al. 2013; Roesch et al. 2012; Teixeira et al. 2010), but focussed mainly on small spatial scales and prokaryotes (but see Lee et al., 2012), while eukaryotes and invertebrates seemed somewhat neglected (but see Dreesens et al., 2014; Niederberger et al., 2015, for studies on fungi). Consequently, subsequent chapters of my work investigated how to increase knowledge regarding the distribution of Antarctic biota both on small and large spatial scales, while enabling insights into the cryptic community-level organisation of terrestrial Antarctic biota.

Chapter 2 Chapter two of this thesis demonstrated how eukaryotic biodiversity information from terrestrial Antarctica can be generated using MHTS. Even the simplest analysis method (network visualisation) revealed substantial eukaryotic biodiversity differences between three sampling regions in the Prince Charles Mountains (Mount Menzies, Mawson Escarpment and Lake Terrasovoje), indicative of a latitudinal or elevational correlation to eukaryotic diversity and richness. Rarefaction methods performed poorly on this Antarctic biological data, due to the low yield of eukaryotic phylotypes from the most extreme environment (Mount Menzies). Principal Coordinate Analysis (PCoA) proved to be a better means to analyse and visualise biodiversity data, particularly to detect trends in highly detailed phylotype data. High-rank taxonomic inspection of phylotypes showed that protists and fungi were most widespread in the sampling area, low-rank taxonomic inspections resulted in the identification of Antarctic fungi (*Cryomycetes antarcticus*), or bird parasites (Apicomplexa: Eimeriidae) where birds (South polar skua - *Catharacta maccormicki*, or Snow petrels - *Pagodroma nivea*) were observed.

Chapter 3 Chapter three of this thesis revealed that fidelities of taxonomic assignments differed depending on metagenetic marker and inspected environment, and application of non-arbitrary sequence processing parameters made it possible to uncover these findings. While it became obvious that morphological description is easier for larger invertebrates, it was shown that *18S rDNA* (18S) is preferable over *cytochrome oxidase subunit I* (COI) for the retrieval of family- and genus-level taxonomic information of Antarctic invertebrates. At the same time, COI performed more accurately in retrieving class- and order-level information. Careful selection of MHTS processing parameters eliminated phylotypes that could not be retrieved in replicated amplifications, and the removal of artefact data, for example chimeras,

reads containing PCR or sequencing errors. Precisely defined phylotypes and comparison to morphological reference data allowed determination of metagenetic marker rank-level fidelity for invertebrate biodiversity assessments.

Chapter 4 In chapter four, MHTS data indicated that distribution of Antarctic invertebrates was constrained by age-related salinity in inland areas. Other environmental variables appeared to play a minor role in constraining biodiversity, but provided for a higher-diverse community in regions with low salinity, such as in coastal proximity. In general, inland areas (Mount Menzies and Mawson Escarpment) appeared to accumulate salt in soils over time, while salts are more likely to be washed out of soils at the wetter, more coastal Lake Terrasovoje or at some inland sites that receive moisture from glacial meltwater (e.g. Magalhaes et al., 2012). This trend was reflected by measurements of S, conductivity, K, NO_3^- , most pronounced at the Mawson Escarpment, and least obvious around Lake Terrasovoje, while Mount Menzies exhibited an intermediate position due to its more recent exposure after glacial recession (Chapter 2, sec. 2.8). Accordingly, Chromadorea (Nematoda) and Monogonata (Bdelloidea) were found distributed across the three main regions, likely due to physiological adaptations to tolerate high salinity and due to usage of saline-tolerant micro-eukaryotic food sources. In low-saline, nutrient-rich areas, phyla Tardigrada, Arachnida and classes Enoplea (Nematoda) and Bdelloidea (Rotifera) dominated.

5.2. High throughput sequencing for Antarctica

Chapter 1 Chapter one showed, in line with the SCAR ANTECO research program (<http://www.scar.org/srp/anteco>) that MHTS needs to be applied with a densely spaced sampling regime over large distances and on a variety of communities to improve terrestrial Antarctic baseline biodiversity information. Substantial effort is currently invested in examining biodiversity and community level interactions of typical Antarctic terrestrial biota, but even many recent publications frequently applied traditional Sanger sequencing technology for this purpose (Altermann et al., 2014; Fernández-Mendoza et al., 2011; Gokul et al., 2013; Jungblut et al., 2012; Khan et al., 2011; Nakai et al., 2012). Practical limitations of early, Sanger-based, metagenetic studies allowed analysis of relatively few samples, even when samples were retrieved over large geographical distances (Fell et al., 2006; Lawley et al., 2004).

The current advent of MHTS approaches in Antarctica so far has mainly appreciated the high taxonomic resolution of the resulting data, but mostly has not sufficiently made use of options to process large sample numbers for all occurring taxa (Bottos et al., 2014; López-Bueno et al., 2009; Makhalanyane et al., 2013; Teixeira et al., 2010). At the same time, climate modellers call for precisely the type of biological information data that could be obtained with HTS platforms (Gutt et al., 2012), also due to the huge spatial heterogeneity of the Antarctic environment (Convey et al., 2014). The Chapter summarizes evidence that the application of MHTS over a large spatial scale with a densely spaced sampling regime would greatly improve baseline biodiversity information from terrestrial Antarctica as shown by recent studies (Lee et al., 2012). The Chapter furthermore supplies background and methodological knowledge to apply such survey methods.

Chapter 2 Co-correlation between overall eukaryotic diversity / richness and elevation / latitude or latitudinal trends observed in Chapter two could indicate a substantially different evolutionary history between the hitherto somewhat neglected Antarctic unicellular eukaryotes and the larger invertebrates. For invertebrates, glacial events and salt accumulation seem to be the main determinants influencing occurrence throughout the continent (Convey and Stevens, 2007; Magalhaes et al., 2012) (also: chapter 4). However, throughout the continent, unicellular eukaryotes are frequently found in old and saline locations, albeit at a lower diversity, where invertebrates were not found (Fell et al., 2006; Magalhaes et al., 2012) (also: chapter 2). Consequently, there currently appears to be no indication that Antarctic unicellular eukaryotes are restricted to younger substrate deposits in the same fashion as invertebrates. Lack of co-correlation between micro-eukaryote biodiversity and latitude was previously attributed to high site isolation and possible endemism of Antarctic eukaryotes (Lawley et al., 2004). Results obtained in this thesis (chapter 2) indicate a much higher diversity of unicellular eukaryotes than observed by Lawley et al. (2004). Also, in line with recent studies (Dreesens et al., 2014; Niederberger et al., 2015) since a much higher diversity was obtained here with MHTS approaches, results indicate that MHTS is better suited to provide biodiversity information from unicellular eukaryotes than early Sanger-based metagenetic studies.

Chapter 3 Findings of Chapter three suggest that the amount of available reference data and MHTS markers for Antarctic taxa should be combined and extended. The

quest to find MHTS markers suitable for eukaryotic taxa also present in Antarctica is on-going (Epp et al., 2012; Hajibabaei et al., 2006; Leray et al., 2013; Machida and Knowlton, 2012; Riaz et al., 2011). At the same time marker-associated reference information may prove to be insufficient for species and genus-level taxonomic assignments of MHTS phylotypes even after decades of data collection (Benson et al., 2011; Folmer et al., 1994), as exemplified here for Antarctic invertebrates and COI. Candidate markers offering comprehensive reference data (here: 18S or COI) for phylotype identification are often criticised for their application in MHTS (Clarke et al., 2014; Deagle et al., 2014; Tang et al., 2012). There may not be a single marker suitable to capture the invertebrate complexity of mixed Antarctic soil extracts, but it was demonstrated here that statistical tests allowed taxonomic fidelity evaluation of combined 18S and COI data using morphological data as a baseline. The fidelity-comparison approach developed in this thesis could be used analogously to evaluate the quality of other MHTS marker candidates, also when applied to other Antarctic taxa. Combinations of different markers could then be chosen based upon their (qualified) ability to retrieve certain taxonomic ranks. Meanwhile, it remains important to increase reference data for MHTS markers such as 18S and COI for research on Antarctic biodiversity.

Chapter 4 In Chapter four, soil salinity and terrain age were found to be the main constraints of invertebrate biodiversity across large spatial scales in the Prince Charles Mountains. These findings presented encompassed a large spatial extent and hence the effect of spatial autocorrelation was mitigated. Additionally, the parallel assessment of four (Chelicerata, Nematoda, Tardigrada, Rotifera) out of five (not Hexapoda) invertebrate (sub-) phyla commonly encountered in Antarctica, was more comprehensive than studies focused on few taxa. Importantly, while most abiotic variables contained independent information to disperse invertebrate classes in the environmental space of the selected Canonical Correspondence Analysis (CCA, Ter Braak, 1986) model in concordance with earlier studies (Courtright et al., 2001), none apart from salinity was found significant in determining invertebrate occurrence, indicating that invertebrates could be able to survive in a variety of substrates, as long as salinity does not exceed physiological tolerance. In summary, results obtained here were indicative for Antarctic-specific terrain-age-related substrate salt accumulation being the main driver of invertebrate biodiversity across inland areas in the Prince Charles Mountains, while in coastal proximity decreased salinity pressure

and higher nutrient input is incidental with a more diverse invertebrate fauna. Since this relationship between salinity and physiological tolerance was inferred from the distribution of invertebrate taxa, future physiological studies will be able to prove or disprove this hypothesis.

5.3. Implications and future improvements

Chapter 1 The relatively simple MHTS workflow applied in this thesis was ideal for large scale approaches, easily automatable using robotics and multiplex pipetting, but may be affected by PCR-associated biases and may be less cost-efficient in comparison to alternative methods. Hybrid capture of DNA in bulk extracts could allow for detection of a broader diversity of homologous gene regions, while reducing PCR-related biases (Denonfoux et al., 2013). A combination of *shotgun sequencing* methods and *amplicon sequencing* could allow for retrieval of, for example, full length COI sequences using HTS technology (Liu et al., 2013). The applied amplicon sequencing methods could be more cost efficient using a modular tagging approach (Clarke et al., 2014)¹. Another option would be the omission of library quantification, while instead pooling libraries by volume (coupled with shearing and reassembly) (Feng et al., 2015). Application of such methodological adjustments could improve read length (i.e. information content) and increase cost effectiveness of the MHTS approaches applied here.

Chapter 2 Despite limitations of rarefaction methods and low sequence coverage, ecological trends from ecological data could be obtained using Principal Coordinate Analysis (PCoA) and the Unifrac distance measure. The usage of proportions or rarefied counts, although applied widely, are inappropriate for analysis of differentially abundant phylotypes in metagenetic data (McMurdie and Holmes, 2014). Furthermore, sequencing depth of metagenetic libraries is crucial for the estimation of reliable biodiversity measures in metagenetic data (Smith and Peay, 2014). R packages *deSeq* and *edgeR* offer alternative ways of phylotype abundance correction (Anders and Huber, 2013; Robinson et al., 2009), and sequencing coverage can be increased by repeated sequencing of the libraries in combination with the already generated data. Combining both approaches will improve retrieval of more reliable biodiversity metrics.

¹See Publications and Awards and appendix

Chapter 3 A newly developed method for comparing metagenetic data to morphological reference data was presented in Chapter 3. Precise low-rank taxonomic assignments with morphological approaches are difficult to achieve for Antarctic invertebrates, such as nematodes (Freckman and Virginia, 1997) or rotifers, tardigrades and arthropods (Powers et al., 1998; Velasco-Castrillón et al., 2014). Furthermore, the developed scoring algorithm (Appendix B) uses unweighted concordance fractions for each taxonomic rank for the determination of intra-class correlation coefficients (ICCs; Koch, 1982), leading to values that are difficult to interpret (Fig. 3.4 and Tab. 3.4). Better morphological reference data will improve marker comparisons; a taxon-rank-weighted concordance algorithm will provide more easily interpretable ICCs.

Chapter 4 The application of the R packages PHYLOSEQ (McMurdie and Holmes, 2013) and VEGAN (Dixon, 2003; Oksanen et al., 2015) to combined MHTS data enabled the application of established, well documented, powerful community analysis methods to large data volumes. The evaluated final CCA model had higher significance than recent comparable approaches that used variance inflation factors rather than Akaike’s information criterion for model selection (Ding et al., 2015). Similar to the analysis described here, forward selection of model constraints have previously been used in other Antarctic studies (Sinclair et al., 2006). Scarcity of biological data is known to impair ecological statistical analysis in Antarctica (Magalhaes et al., 2012), and consequently non-agglomerated phylotype data was difficult to analyse with various ordination methods such as *multidimensional scaling* (MDS) (Wish and Carroll, 1982), *constrained principal component analysis* (CAP) and *redundancy analysis* (RDA) (Legendre and Andersson, 1999) due to the low spatial overlap of individual phylotypes. Higher significance values for environmental observations could have been achieved by inspecting variables individually as shown elsewhere (Sinclair et al., 2006). Analysis of non-agglomerated phylotype abundance values would have been possible with Hellinger transformation of abundance values (Legendre, 2008). Hellinger-transformed phylotype data could have also been used for weighted PCA (i.e. CAP) accounting for spatial autocorrelation (Caruso et al., 2010). Finally, effects of marker specific phylotype abundance biases, here mitigated through agglomeration, could have been addressed using more recent abundance correction methods mentioned above (Anders and Huber, 2013; Robinson et al., 2009). Sampling design, taxon agglomeration and application of CCA accounted for

potential pitfalls of the analyses performed here, such as effects of spatial autocorrelation, marker specific abundance biases, and gradient distortion related to the latter two effects (Caruso et al., 2010; Palmer, 1993); however, an extensive set of tools and analysis methods are available to further detail findings obtained in the future.

5.4. Conclusion

Current Antarctic biology is majorly influenced by the desire to conserve the unique and still largely uncharacterised biodiversity of the continent and surrounding islands. Of major concern are the establishment of non-indigenous species in Antarctica's protected regions that threaten indigenous species (McGeoch et al., 2015; Shaw et al., 2014). Plants (e.g. *Poa annua*) and insects (e.g. *Rhopalosiphum padi*) make up the vast majority of alien species in the Antarctic region and are most widespread with close to half of all bioregions occupied (McGeoch et al., 2015). At the same time, missing baseline information of endemic and invasive taxa hinders the assessment of effectiveness of conservation measurements (McGeoch et al., 2015). Also, definition and extension of protected regions in Antarctica, particularly in remote locations, is urgently required (Shaw et al., 2014), and inevitably coupled with obtaining baseline biological data from those remote regions. Systematic measurements of the impact of alien species will require a systematic network to conserve Antarctic biodiversity as a whole (McGeoch et al., 2015; Shaw et al., 2014).

Highlighted in this thesis were the possibilities of MHTS approaches to (a) survey for small and cryptic eukaryotic organisms in remote and hitherto largely uncharacterised regions of Antarctica such as the Prince Charles Mountains, (b) detect taxa of all major invertebrate (sub-) phyla encountered in Antarctica with relatively simple, fast, and standardised methods across large spatial scales, (c) generate complementary biodiversity information using multiple MHTS markers, and (d) relating invertebrate occurrences to environmental constraints with the help of rich geo-referenced observational data.

Such methods are only now being recognised for their ability to monitor for the distinction and occurrence of invasive taxa in Antarctica, but seemingly are not applied widely to achieve this goal (Chown et al., 2015). Consequently, methods developed and applied here should be furthered, for example, to define and extend protected regions in Antarctica, particularly in remote regions (Shaw et al., 2014; Terauds et al., 2012). More specifically, methods presented here could inform on

the number of eukaryotic alien and invasive species (compared to endemic) per biogeographic region in the DPSR framework (McGeoch et al., 2015), coupled with the collection baseline biodiversity information from hitherto unsurveyed regions.

Requirements to enable large scale biomonitoring across Antarctica using MHTS are availability, linking and usability of reference sequence data, biological records and distributional data (Chown et al., 2015). The work presented here provides examples for generation and usage of MHTS sequence information from remote Antarctic habitats, demonstrates how information retrieved using different metagenetic markers can be combined, provides methods for quality assessment of different MHTS markers and finally shows the application of combined information to investigate current Antarctic environmental constraints on continental fauna. Therefore, this thesis serves as a valuable resource to formalise and organise large-scale biodiversity information from Antarctic terrestrial habitats, in line with the latest findings (Chown et al., 2015).

5.5. Acknowledgements

I thank Mark Stevens (South Australian Museum) and Laurence Clarke (Australian Antarctic Division) for helpful comments on the manuscript.

A. Phylotype information chapter 2

Results of non-parametric ANOVA are shown in the in following table. Provided are phylotype abundances in groups of soil samples. Only eukaryotic phylotypes that contributed by at least one percent to the total abundance per sample were included into analysis, “n” describes the number of samples across which the the phylotype was found. No phylotype was found in more then six soil samples. All statistic values rounded from 13 decimal digits. “Phylotype” - phylotype descriptor, “*p*” - uncorrected *p* value, “FRR - *p*” - false discovery rate corrected *p* value, “Bonferroni *p*” - Bonferroni corrected *p* value. Bonferroni corrected *p* values below 0.05 marked as significant with “*”. Taxonomy assignment performed in QIIME against the SILVA database, release 111. “Similarity” - Similarity of phylotype sequence cluster to query sequence during taxonomy assignment, calculated with UCLUST.

n	Phylotype	Test-Statistic	<i>p</i>	FDR <i>p</i>	Bonferroni <i>p</i>	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
6	denovo7773	6.8804	0.032	0.0320	0.0320 *	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes; Heterodermia; Heterodermia boryi	96.9 %
5	denovo4100	8.3695	0.015	0.0305	0.0305 *	67.5	900.25	0.0	Eukaryota; Stramenopiles; Peronosporomycetes; Phytium; uncultured Eimeriidae	96.7 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
5	denovo7773	6.8804	0.032	0.0321	0.0641	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes;	96.9 %
4	denovo3376	10.4554	0.005	0.0125	0.0376 *	0.0	1520.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured cercozoan	97.7 %
4	denovo7101	10.4554	0.005	0.0125	0.0376 *	0.0	942.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured eukaryote	98.4 %
4	denovo7803	10.4554	0.005	0.0125	0.0376 *	0.0	957.0	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	90.7 %
4	denovo4100	8.3696	0.015	0.0266	0.1066	67.5	900.25	0.0	Eukaryota; Stramenopiles; Peronosporomycetes; Phytium; uncultured Eimeriidae	96.7 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
4	denovo7773	6.8805	0.0321	0.0449	0.2244	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes; Heterodermia; Heterodermia boryi	96.9 %
4	denovo3058	5.6770	0.0585	0.0683	0.4096	74.25	647.5	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	98.4 %
4	denovo4853	1.4839	0.4762	0.4762	1.0	148.5	825.0	0.25	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; uncultured eukaryote	98.1 %
3	denovo3376	10.4554	0.0054	0.0358	0.1073	0.0	1520.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured; uncultured cerczoan	97.7 %
3	denovo7101	10.4554	0.0054	0.0358	0.1073	0.0	942.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured; uncultured eukaryote	98.4 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
3	denovo7803	10.4554	0.0054	0.0358	0.1073	0.0	957.0	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	90.7 %
3	denovo4100	8.3696	0.0152	0.0508	0.3045	67.5	900.25	0.0	Eukaryota; Stramenopiles; Peronosporomycetes; Phytium; uncultured Eimeriidae	96.7 %
3	denovo931	7.1566	0.0279	0.0508	0.5585	0.0	868.0	0.0	Eukaryota; Opisthokonta; Fungi; Basal fungi; Zoopagomycotina; Plakinidae sp.	96.3 %
3	denovo1454	7.1566	0.0279	0.0508	0.5585	0.0	3090.0	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Phyllopharyngea; Cyrtophoria; Chilodonella; uncultured microeukaryote	93.8 %
3	denovo2978	7.1566	0.0279	0.0508	0.5585	0.0	1332.25	0.0	Eukaryota; Opisthokonta; Metazoa; Arthropoda; Chelicerata; Arachnida; Microcaeculus; Microcaeculus sp.	92.4 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
3	denovo3630	7.1566	0.0279	0.0508	0.5585	0.0	1194.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; Pelagodileptus; Dileptus sp.	98.1 %
3	denovo4147	7.1566	0.0279	0.0508	0.5585	0.0	1595.5	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Heterococcus; Heterococcus sp.	99.2 %
3	denovo5400	7.1566	0.0279	0.0508	0.5585	0.0	708.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Teratosphaeriaceae; Staninwardia; Staninwardia suttonii	98.4 %
3	denovo5497	7.15662	0.0279	0.0508	0.5585	0.0	1032.5	0.0	Eukaryota; Amoebozoa; Conosa; Variosea; Varipodida; Flamella; uncultured Eimeriidae	100 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
3	denovo7773	6.8805	0.0321	0.0534	0.6411	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes; Heterodermia; Heterodermia boryi	96.9 %
3	denovo3058	5.6770	0.0585	0.0900	1.0	74.25	647.5	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	98.4 %
3	denovo4797	2.8493	0.2406	0.3007	1.0	0.0	1133.25	65.25	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Agaricomycotina; Tremellomycetes; Syzygospora; Syzygospora bachmannii	95.3 %
3	denovo6120	2.8494	0.2406	0.3007	1.0	88.5	656.0	0.0	Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Trebouxioophyceae; Xylochloris; Xylochloris irregularis	92.2 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
3	denovo7292	2.8494	0.2406	0.3007	1.0	0.25	1163.0	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes; Phaeoacremonium; Phaeoacremonium rubrigenum	96.1 %
3	denovo589	2.3855	0.3034	0.3193	1.0	2820.75	98.75	0.0	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Pucciniomycotina; Microbotryomycetes; Rhodosporidium; Rhodosporidium fluviale	95.3 %
3	denovo1569	2.3855	0.3034	0.3193	1.0	768.0	61.5	0.0	Eukaryota; Rhizaria; Cerczoa; Cercomonadidae; Cavernomonas; Cavernomonas stercoris	91.8 %
3	denovo3089	2.3855	0.3034	0.3193	1.0	716.0	59.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Nassophorea; Obertrumia; uncultured eukaryote	100 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
3	denovo4853	1.4839	0.4762	0.4762	1.0	148.5	825.0	0.25	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; uncultured eukaryote	98.1 %
2	denovo3376	10.4554	0.0053657303905	0.0465	0.1395	0.0	1520.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured; uncultured cerczoan	97.7 %
2	denovo7101	10.4554	0.0053657303905	0.0465	0.1395	0.0	942.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured; uncultured eukaryote	98.4 %
2	denovo7803	10.4554	0.0053657303905	0.0465	0.1395	0.0	957.0	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	90.7 %
2	denovo4100	8.36957	0.0152255154768	0.0660	0.3959	67.5	900.25	0.0	Eukaryota; Stramenopiles; Peronosporomycetes; Phytium; uncultured Eimeriidae	96.7 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo931	7.15663	0.0279227571824	0.0660	0.7260	0.0	868.0	0.0	Eukaryota; Opisthokonta; Fungi; Basal fungi; Zoopagomycotina; uncultured Plakinidae sp.	96.3 %
2	denovo1454	7.15663	0.0279227571824	0.0660	0.7260	0.0	3090.0	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Phyllopharyngea; Cyrtophoria; Chilodonella; uncultured microeukaryote	93.8 %
2	denovo2978	7.15663	0.0279227571824	0.0660	0.7260	0.0	1332.25	0.0	Eukaryota; Opisthokonta; Metazoa; Arthropoda; Chelicerata; Arachnida; Microcaeculus; Microcaeculus sp.	92.4 %
2	denovo3630	7.15663	0.0279227571824	0.0660	0.7260	0.0	1194.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; Pelagodileptus; uncultured Dileptus	98.1 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo4147	7.15663	0.0279227571824	0.0660	0.7260	0.0	1595.5	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Heterococcus; Heterococcus sp. W1232	99.2 %
2	denovo5400	7.15663	0.0279227571824	0.0660	0.7260	0.0	708.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Teratosphaeriaceae; Staninwardia; Staninwardia suttonii	98.4 %
2	denovo5497	7.15663	0.0279227571824	0.0660	0.7260	0.0	1032.5	0.0	Eukaryota; Amoebozoa; Conosa; Variosea; Varipodida; Flamella; uncultured Eimeriidae	100 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo7773	6.88048	0.0320570213789	0.0660	0.8335	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes; Heterodermia; Heterodermia boryi	96.9 %
2	denovo3058	5.67698	0.0585139495684	0.1170	1.0	74.25	647.5	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	98.4 %
2	denovo1568	4.36364	0.112836187317	0.1544	1.0	987.25	0.0	0.0	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Agaricomycotina; Agaricomycetes; Cryptococcus; Cryptococcus huempfi	93.0 %
2	denovo2186	4.36364	0.112836187317	0.1544	1.0	0.0	1862.0	0.0	Eukaryota; Rhizaria; Cerczoa; Silicoflosea; Euglyphida; Euglyphidae; Euglypha; Euglypha rotunda	94.6 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo2399	4.36364	0.112836187317	0.1544	1.0	0.0	702.25	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Xanthophyceae sp.	92.8 %
2	denovo2451	4.36364	0.112836187317	0.1544	1.0	0.0	846.0	0.0	Eukaryota; Rhizaria; Cercozoa; Thecofilosea; Cryomonadida; Protaspidae; Protaspa; uncultured eukaryote	90.6 %
2	denovo6174	4.36364	0.112836187317	0.1544	1.0	0.0	857.25	0.0	Eukaryota; Amoebozoa; uncultured Eimeriidae	91.8 %
2	denovo7249	4.36364	0.112836187317	0.1544	1.0	0.0	4241.5	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Colpodea; Platyophryida; Platyophrya; uncultured eukaryote	91.7 %
2	denovo4797	2.84940	0.240580916428	0.284322901233	1.0	0.0	1133.25	65.25	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Agaricomycotina; Tremellomycetes; Syzygospora; Syzygospora bachmannii	95.3 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo6120	2.84940	0.240580916428	0.2843	1.0	88.5	656.0	0.0	Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Trebouxiophyceae; Xylochloris; Xylochloris irregularis	92.2 %
2	denovo7292	2.84940	0.240580916428	0.2843	1.0	0.25	1163.0	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes; Phaeoacremonium; Phaeoacremonium rubrigenum	96.1 %
2	denovo589	2.38554	0.303379408252	0.315514584582	1.0	2820.75	98.75	0.0	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Pucciniomycotina; Microbotryomycetes; Rhodosporidium; Rhodosporidium fluviale	95.3 %
2	denovo1569	2.38554	0.303379408252	0.3155	1.0	768.0	61.5	0.0	Eukaryota; Rhizaria; Cercozoa; Cercomonadidae; Cavernomonas; Cavernomonas stercoris	91.8 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
2	denovo3089	2.38554	0.303379408252	0.3155	1.0	716.0	59.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Nassophorea; Obertrumia; uncultured eukaryote	100 %
2	denovo4853	1.48391	0.476181856833	0.4762	1.0	148.5	825.0	0.25	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; uncultured eukaryote	98.1 %
1	denovo7803	12.0038	0.00288622355758	0.0837	0.0837	0.0	957.0	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	90.7 %
1	denovo931	8.67222	0.00796141718188	0.1154	0.2309	0.0	868.0	0.0	Eukaryota; Opisthokonta; Fungi; Basal fungi; Zoopagomycotina; uncultured Plakinidae sp.	96.3 %
1	denovo3058	6.07521	0.0213872660897	0.1628	0.6202	74.25	647.5	0.0	Eukaryota; Rhizaria; Cerczoa; Glissomonadida; uncultured eukaryote	98.4 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo4100	5.96145	0.0224539628653	0.1628	0.6512	67.5	900.25	0.0	Eukaryota; Stramenopiles; Peronosporomycetes; Phytium; uncultured Eimeriidae	96.7 %
1	denovo4147	4.41216	0.0461883154958	0.2679	1.0	0.0	1595.5	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Heterococcus; Heterococcus sp.	99.2 %
1	denovo2978	3.98512	0.0576095945169	0.2784	1.0	0.0	1332.25	0.0	Eukaryota; Opisthokonta; Metazoa; Arthropoda; Chelicerata; Arachnida; Microcaeculus; Microcaeculus sp.	92.4 %
1	denovo5497	2.76664	0.115732300704	0.4335	1.0	0.0	1032.5	0.0	Eukaryota; Amoebozoa; Conosa; Variosea; Varipodida; Flamella; uncultured Eimeriidae	100 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo7773	2.32407	0.153553518506	0.4335	1.0	79.0	2340.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Lecanoromycetes; Heterodermia; Heterodermia boryi	96.9 %
1	denovo3630	2.25686	0.16054720588	0.4335	1.0	0.0	1194.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; Pelagodileptus; uncultured Dileptus	98.1 %
1	denovo5400	1.52748	0.268439164246	0.4335	1.0	0.0	708.5	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Teratosphaeriaceae; Staninwardia; Staninwardia suttonii	98.4 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo2186	1.41390	0.292429400037	0.4335	1.0	0.0	1862.0	0.0	Eukaryota; Rhizaria; Cerczoa; Silicoflosea; Euglyphida; Euglyphidae; Euglypha; Euglypha rotunda	94.6 %
1	denovo4797	1.40038	0.295459367549	0.4335	1.0	0.0	1133.25	65.25	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Agaricomycotina; Tremellomycetes; Syzygospora; Syzygospora bachmannii	95.3 %
1	denovo7101	1.39194	0.297366815833	0.4335	1.0	0.0	942.75	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; uncultured; uncultured eukaryote	98.4 %
1	denovo1454	1.38093	0.29988062543	0.4335	1.0	0.0	3090.0	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Phyllopharyngea; Cyrtophoria; Chilodonella; uncultured microeukaryote	93.8 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo4853	1.35665	0.305515732089	0.4335	1.0	148.5	825.0	0.25	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Litostomatea; Haptoria; uncultured eukaryote	98.1 %
1	denovo6120	1.32361	0.313394982342	0.4335	1.0	88.5	656.0	0.0	Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Trebouxioephyceae; Xylochloris; Xylochloris irregularis	92.2 %
1	denovo3376	1.25748	0.329921234262	0.4335	1.0	0.0	1520.75	0.0	Eukaryota; Rhizaria; Cercospora; Thecofilosea; uncultured; uncultured cercosporan	97.7 %
1	denovo6174	1.24995	0.331869268709	0.4335	1.0	0.0	857.25	0.0	Eukaryota; Amoebozoa; uncultured Eimeriidae	91.8 %
1	denovo2399	1.10078	0.37354072817	0.4335	1.0	0.0	702.25	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Xanthophyceae sp.	92.8 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo7249	1.04094	0.392041945338	0.4335	1.0	0.0	4241.5	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Colpodea; Platyophryda; Platyophrya; uncultured eukaryote	91.7 %
1	denovo7292	1.00438	0.403893988351	0.4335	1.0	0.25	1163.0	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes; Phaeoacremonium; Phaeoacremonium rubrigenum	96.1 %
1	denovo2451	1.00079	0.405083060065	0.4335	1.0	0.0	846.0	0.0	Eukaryota; Rhizaria; Cerczoa; Thecofilosea; Cryomonadida; Protaspidae; Protaspa; uncultured eukaryote	90.6 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo1568	1.00068	0.405120458466	0.4335	1.0	987.25	0.0	0.0	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Agaricomycotina; Agaricomycetes; Cryptococcus; Cryptococcus huempii	93.0 %
1	denovo1826	1.0	0.405344429706	0.4335	1.0	0.0	5733.75	0.0	Eukaryota; Stramenopiles; Xanthophyceae; Tribonematales; Pseudopleurochloris; Pseudopleurochloris antarctica	91.2 %
1	denovo3671	1.0	0.405344429706	0.4335	1.0	0.0	985.75	0.0	Eukaryota; Opisthokonta; Fungi; Ascomycota; Pezizomycotina; Leotiomycetes; Chlorociboria; Chlorociboria aeruginosa	98.4 %

n	Phylotype	Test-Statistic	p	FDR p	Bonferroni p	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo4939	1.0	0.405344429706	0.4335	1.0	0.0	1480.25	0.0	Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Trebouxioiphyceae; Parachlorella; Mucidosphaerium; Mucidosphaerium sphagnale	92.2 %
1	denovo589	0.9657	0.416911398451	0.4335	1.0	2820.75	98.75	0.0	Eukaryota; Opisthokonta; Fungi; Basidiomycota; Pucciniomycotina; Microbotryomycetes; Rhodosporidium; Rhodosporidium fluviale	95.3 %
1	denovo1569	0.9212	0.432531112156	0.4335	1.0	768.0	61.5	0.0	Eukaryota; Rhizaria; Cerczoa; Cercomonadidae; Cavernomonas; Cavernomonas stercoris	91.8 %

n	Phylotype	Test-Statistic	<i>p</i>	FDR <i>p</i>	Bonferroni <i>p</i>	\bar{x} ME	\bar{x} LT	\bar{x} MM	Taxonomy	Similarity
1	denovo3089	0.9187	0.433453501576	0.4335	1.0	716.0	59.25	0.0	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Conthreep; Nassophorea; Obertrumia; uncultured eukaryote	100 %

B. Analysis code chapter 3

Saved: 2/06/15 15:37:32

```

1 ## 15.04.29 - Re-analysis for Chapter 3 - Shell part
2 ## =====
3 # modified from /Users/paul/Documents/140911_c3_analysis/1_analysis_shell_COI-18S-AVC/150302_C3_metagenetic_approach.txt
4 # see R part: /Users/paul/Documents/140911_c3_analysis/2_analysis_R_scripts/150504_C3_revisions.r
5
6 ## Step 1: Get Invertebrate OTU tables that are not filtered on abundance
7 ## =====
8 ## This step is okay
9 # filter for inverts
10 # copy to local
11
12 ## 18S:
13 ## ----
14 # find all Metazoans
15 target="/mnt/paul_folder/140730_18S_data/141211_18S_DeNovo/141211_18S_tax_assignment_90/141211_18S_rep_set_099_vs_db_100/141211_rep_set_099_tax_assignments.txt"
16 awk -F '\t;' ' $4 == " __Metazoa" {print $5,$6 }' "$target" | sort -d | uniq -c
17 # create output directory
18 targetdir="/mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates"
19 mkdir -p $targetdir
20 # check input file paths - choose only files with metadata and taxonomy assignments
21 ls /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/*.biom
22 # load qiime
23 module load Qiime/1.8.0
24 # loop
25 for file in /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/*.biom; do
26 [[ -e "$file" ]] || continue
27 echo @ processed file: "$file"
28 out_tmp1=$targetdir/${basename $file .biom}.tmp1
29 echo @ temp file 1 "$out_tmp1"
30 out_tmp2=$targetdir/${basename $file .biom}.tmp2
31 echo @ temp file 2 "$out_tmp2"
32 out_biom="$targetdir"/${basename "$file" .biom}_invertebrates.biom
33 echo @ biom file 2 "$out_biom"
34 echo @ retaining invertebrates
35 # filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p __Nematoda,__Rotifera,__Tardigrada,__Chelicerata
36 filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p __Nematoda,__Rotifera,__Tardigrada,__Arthropoda
37 echo @ filter empty samples
38 filter_samples_from_otu_table.py -i "$out_tmp1" -o "$out_tmp2" -n 1
39 echo @ filter 5 count OTUs
40 filter_otus_from_otu_table.py -i "$out_tmp2" -o "$out_biom" -n 5
41 echo @ generating summary files
42 biom summarize-table -i "$out_biom" -o "$targetdir"/${basename "$out_biom" .biom}.sum_qual.txt --qualitative
43 biom summarize-table -i "$out_biom" -o "$targetdir"/${basename "$out_biom" .biom}.sum_quant.txt
44 echo @ erasing temp files
45 rm -v /mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates/*.tmp?
46 done
47
48 ## COI:
49 ## ----
50 # create output directory
51 targetdir="/mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates"

```

Saved: 2/06/15 15:37:32

```

52 mkdir -p $targetdir
53 # check input file paths - choose only files with metadata and taxonomy assignments
54 ls /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/*md_assigned_only.biom
55 # load qiime
56 module load Qiime/1.8.0
57 # loop
58 for file in /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/*md_assigned_only.biom; do
59     [[ -e "$file" ]] || continue
60     echo @ processed file: "$file"
61     out_tmp1=$targetdir/${basename $file .biom}.tmp1
62     echo @ temp file 1 "$out_tmp1"
63     out_tmp2=$targetdir/${basename $file .biom}.tmp2
64     echo @ temp file 2 "$out_tmp2"
65     out_biom="$targetdir"/${basename "$file" .biom}_invertebrates.biom
66     echo @ biom file 2 "$out_biom"
67     echo @ retaining invertebrates
68     # filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p Chelicerata,Nematoda,Rotifera,Tardigrada
69     filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p Arthropoda,Nematoda,Rotifera,Tardigrada
70     echo @ filter empty samples
71     filter_samples_from_otu_table.py -i "$out_tmp1" -o "$out_tmp2" -n 1
72     echo @ filter 5 count OTUs
73     filter_otus_from_otu_table.py -i "$out_tmp2" -o "$out_biom" -n 5
74     echo @ generating summary files # update pathnames
75     biom summarize-table -i "$out_biom" -o "$targetdir"/${basename "$out_biom" .biom}.sum_qual.txt --qualitative
76     biom summarize-table -i "$out_biom" -o "$targetdir"/${basename "$out_biom" .biom}.sum_quant.txt
77     echo @ erasing temp files
78     rm -v /mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates/*.tmp?
79 done
80
81 ## copy to local
82 # 18S:
83 local_file="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150302_18S"
84 remote_file="/mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates/*"
85 scp -r -C -i /Users/paul/Documents/06_shell/acad_keypair.pem pczechowski@130.220.209.109:"$remote_file" "$local_file"
86 # COI:
87 local_file="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150302_COI"
88 remote_file="/mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates/*"
89 scp -r -C -i /Users/paul/Documents/06_shell/acad_keypair.pem pczechowski@130.220.209.109:"$remote_file" "$local_file"
90
91
92 ## Step 2: filter by different abundances and sort data
93 ## =====
94 ## This step is okay
95
96 # filtering percentages
97 percentage[1]="0.001"
98 percentage[2]="0.002"
99 percentage[3]="0.003"
100 percentage[4]="0.005"
101 # input file lists
102 list[1]="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150302_18S/*.biom"

```

```
103 list[2]="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150302_COI/*.biom"
104 # mapping files
105 mf[1]="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_18S_MF.txt"
106 mf[2]="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_COI_MF.txt"
107 # define path and filenames
108 outdir[1]="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_18S_abundance"
109 outdir[2]="/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_abundance"
110 # parameter file
111 pf="/Users/paul/Documents/140911_c3_analysis/2_analysis_otime_PFs/150215_plot_pf.txt"
112 # loop over percentages
113 for ((k = 1; k <= "${#percentages[@}"; k++)); do
114     echo "@enter loop filtering out fraction: ${percentage[k]}"
115     # loop over lists
116     for ((j = 1; j <= "${#list[@}"; j++)); do
117         echo "@enter loop list: ${list[j]}"
118         # loop over files
119         for file in ${list[j]}; do
120             [[ -e $file ]] || continue
121             echo "@enter loop table: $file"
122             # filter by abundance
123             abdir="${outdir[j]}/${150304}_${percentage[k]}_discarded && mkdir -p "$abdir"
124             # @ defined and created abundance directory: $abdir"
125             abundance="$abdir"/$(basename "$file" .biom).tmp
126             echo "@ defined filename: $abundance"
127             echo "@ retaining percentage"
128             filter_otus_from_otu_table.py --min_count_fraction "${percentage[k]}" -i "$file" -o "$abundance"
129             # retain controls in single file
130             antarctic="$abdir"/$(basename "$abundance" .tmp)_ANT_ "${percentage[k]}.tmp
131             echo "@ defined filename: antarctic"
132             australia="$abdir"/$(basename "$abundance" .tmp)_AUST_ "${percentage[k]}.tmp
133             echo "@ defined filename: australia"
134             all_data="$abdir"/$(basename "$abundance" .tmp)_ALL_ "${percentage[k]}.tmp
135             echo "@ defined filename: $all_data"
136             echo "@ retaining Antarctic and Australian controls"
137             filter_samples_from_otu_table.py -i "$abundance" -o "$antarctic" -m "${mf[j]}" -s "Location:E_Ant_coast"
138             filter_samples_from_otu_table.py -i "$abundance" -o "$australia" -m "${mf[j]}" -s "Location:Australia"
139             merge_otu_tables.py -i "$australia", "$antarctic" -o "$all_data"
140             # isolate soil controls from Australian samples
141             aust_soil="$abdir"/$(basename "$australia" .tmp)_SOIL.tmp
142             echo "@ defined filename: aust_soil"
143             echo "@ isolating Australian soil controls"
144             filter_samples_from_otu_table.py -i "$australia" -o "$aust_soil" -m "${mf[j]}" -s "XtrOri:XTR_pos"
145             # isolate mock controls from Australian samples
146             aust_mock="$abdir"/$(basename "$australia" .tmp)_MOCK.tmp
147             echo "@ defined filename: aust_mock"
148             echo "@ isolating Australian mock controls"
149             filter_samples_from_otu_table.py -i "$australia" -o "$aust_mock" -m "${mf[j]}" -s "XtrOri:PCR_pos"
150             # defining output filenames
151             all_data_b="$abdir"/$(basename "$all_data" .tmp).biom
152             echo "@ defined output filename: $all_data_b"
153             antarctic_b="$abdir"/$(basename "$antarctic" .tmp).biom
```

Saved: 2/06/15 15:37:32

```

154 echo "@ defined output filename: $antarctic_b"
155 aust_soil_b="$abdir"/$(basename "$aust_soil" .tmp).biom
156 echo "@ defined output filename: $aust_soil_b"
157 aust_mock_b="$abdir"/$(basename "$aust_mock" .tmp).biom
158 echo "@ defined output filename: $aust_mock_b"
159 # final filtering
160 echo "@ removing 0 count OTUs"
161 filter_otus_from_otu_table.py -i "$all_data" -o "$all_data_b" -n 1
162 filter_otus_from_otu_table.py -i "$antarctic" -o "$antarctic_b" -n 1
163 filter_otus_from_otu_table.py -i "$aust_soil" -o "$aust_soil_b" -n 1
164 filter_otus_from_otu_table.py -i "$aust_mock" -o "$aust_mock_b" -n 1
165 # convert to text
166 echo "@ converting to text"
167 all_data_txt="$abdir"/$(basename "$all_data_b" .biom).txt
168 antarctic_txt="$abdir"/$(basename "$antarctic_b" .biom).txt
169 aust_soil_txt="$abdir"/$(basename "$aust_soil_b" .biom).txt
170 aust_mock_txt="$abdir"/$(basename "$aust_mock_b" .biom).txt
171 biom convert -i "$all_data_b" -o "$all_data_txt" -b --header-key taxonomy --output-metadata-id "Taxonomy"
172 biom convert -i "$antarctic_b" -o "$antarctic_txt" -b --header-key taxonomy --output-metadata-id "Taxonomy"
173 biom convert -i "$aust_soil_b" -o "$aust_soil_txt" -b --header-key taxonomy --output-metadata-id "Taxonomy"
174 biom convert -i "$aust_mock" -o "$aust_mock_txt" -b --header-key taxonomy --output-metadata-id "Taxonomy"
175 # summaries
176 echo "@ getting summaries"
177 biom summarize-table -i "$all_data_b" -o "$abdir"/$(basename "$all_data_b" .biom).sum_qual.txt --qualitative
178 biom summarize-table -i "$all_data_b" -o "$abdir"/$(basename "$all_data_b" .biom).sum_quant.txt
179 biom summarize-table -i "$antarctic_b" -o "$abdir"/$(basename "$antarctic_b" .biom).sum_qual.txt --qualitative
180 biom summarize-table -i "$antarctic_b" -o "$abdir"/$(basename "$antarctic_b" .biom).sum_quant.txt
181 biom summarize-table -i "$aust_soil_b" -o "$abdir"/$(basename "$aust_soil_b" .biom).sum_qual.txt --qualitative
182 biom summarize-table -i "$aust_soil_b" -o "$abdir"/$(basename "$aust_soil_b" .biom).sum_quant.txt
183 biom summarize-table -i "$aust_mock_b" -o "$abdir"/$(basename "$aust_mock_b" .biom).sum_qual.txt --qualitative
184 biom summarize-table -i "$aust_mock_b" -o "$abdir"/$(basename "$aust_mock_b" .biom).sum_quant.txt
185 # erasing temp files
186 echo "@ erasing temp files"
187 if [ -d "$abdir" ]; then
188 rm -v "$abdir"/*.*.tmp
189 fi
190
191 done
192 done
193
194 ## Step 3: Do all other analyses
195 ## =====
196
197 ## R script at
198 ## /Users/paul/Documents/140911_c3_analysis/2_analysis_R_scripts/150504_C3_revisions.r
199

```

```
1 ## 15.04.29 - Re-analysis for Chapter 3 - R part
2 ## =====
3 ## see shell part /Users/paul/Documents/140911_c3_analysis/1_analysis_shell_COI-18S-AVC/150429_c3_metagenetic_approach.txt
4
5 ### Chapter 3 analysis:
6 # 1: Import all biom files derived from soil data
7
8 # - Evaluation of Parameters:
9 # 2: Plot effect of filtering on 18S and COI soil data
10 # 3: Get tables with highest numbers of Phylotype counts into new list
11 # 4: Get Similarity values between replicates of two soil controls
12 # 5: Combine count and similarity values
13 # 6: * Create display Items: Plot Phylotype counts and Similarity between replicates
14
15 # - Evaluation of Insect controls:
16 # 7: Generate comparale objects of insect controls for plotting
17 # 7a: Plot composition of "Artificial blends" reference data
18 # 8: Compare similarities between insect reference and insect controls
19 # 9: * Create display Items: Heatmap as comparsion between samples
20 # 10: * Create display Items: Barplot as comparsion between samples
21
22 # - Evaluation of Antarctic samples:
23 # 11: Import Antarctic morphology data and metagenetic samples
24 # 11a: Plot composition of "Antarctic soil" reference data
25 # 12: compare assignemnts
26 # 13: plot taxonomic composition
27 # 14: Format Antarctic data into a more useful data frame
28
29 # - Stats
30
31 # 15: Calculate ICC on Control and Antarctic data
32
33 # 99: trials
34
35
36 ### clear environment, set working directory
37 {
38   rm(list=ls()) # clear R environment
39   setwd("/Users/paul/Documents/140911_c3_analysis/2_analysis_R_scripts/") # working directory is here
40
41
42 ### load packages
43 require("phyloseq")
44 require("biom")
45 require("plyr")
46 require("dplyr")
47 require("vegan")
48 require("gplots")
49 require("ade4")
50 require("ggplot2")
51 require("reshape2")
52 require("gridExtra")
```



```

53 require("foreach")
54 require("irr")
55 }
56
57 ##### 1: data import
58 {
59 # Find all .biom files and store in lists, combine list, name list itmes
60 files_18S <- list.files("/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_18S_soil/", full.names=TRUE, pattern="*.biom")
61 files_COI <- list.files("/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/", full.names=TRUE, pattern="*.biom")
62 #file_list <- append(files_18S,files_COI)
63 #names(file_list) <- file_list[] # filenames become names of list items
64 names(files_18S) <- files_18S[] # filenames become names of list items
65 names(files_COI) <- files_COI[] # filenames become names of list items
66
67 # Make a function to process each file
68 # needs file list, returns biom objects in list
69 createPhyloseq <- function(list_item) {
70   # import biom file with pathname in list
71   physeq_ob <- try(import_biom(list_item))
72   # return biom object
73   return(physeq_ob)
74 }
75
76 # read in biom files and generate a list of Phyloseq objects
77 psob_list_18S <- lapply(files_18S, createPhyloseq)
78 psob_list_COI <- lapply(files_COI, createPhyloseq)
79
80 # remove failed imports form list
81 psob_list_COI[which(names(psob_list_COI) %in% c("/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust97_tassgn75_md_assign
82 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust97_tassgn80_md_assigned_only_invertebrates_AUST_0.002_SOIL.bi
83 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust97_tassgn80_md_assigned_only_invertebrates_AUST_0.003_SOIL.bi
84 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust97_tassgn80_md_assigned_only_invertebrates_AUST_0.005_SOIL.bi
85 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust97_tassgn90_md_assigned_only_invertebrates_AUST_0.001_SOIL.bi
86 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust99_tassgn75_md_assigned_only_invertebrates_AUST_0.003_SOIL.bi
87 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust99_tassgn75_md_assigned_only_invertebrates_AUST_0.005_SOIL.bi
88 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust99_tassgn80_md_assigned_only_invertebrates_AUST_0.002_SOIL.bi
89 "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_soil/150121_COI_OTUs_clust99_tassgn90_md_assigned_only_invertebrates_AUST_0.001_SOIL.bi
90
91 # finished data - lists of PhyloSeq objects:
92 psob_list_18S
93 psob_list_COI
94 }
95
96 ##### 2: Plot effect of filtering
97 {
98 # get phylotype count in each table and store in integer (vector?)
99 ets_counts <- sapply(psob_list_18S, function(x){nrow(otu_table(x))})
100 coi_counts <- sapply(psob_list_COI, function(x){nrow(otu_table(x))})
101
102 # function: simplify the names of each element - 18S
103 ShortenNames18S <- function(ct) {
104   names(ct) <- gsub("_md_assigned_only_no_contaminants_invertebrates_", "_", substr(names(ct), 87, 168))

```

```

105 names(ct) <- gsub("_OTUs_clust", "S_c", names(ct))
106 names(ct) <- gsub("_tassgn", "t", names(ct))
107 names(ct) <- gsub("_tassgn", "t", names(ct))
108 names(ct) <- gsub("_AUST_", "f", names(ct))
109 return(ct)
110 }
111
112 # function: simplify the names of each element - COI
113 ShortenNamesCOI <- function(ct) {
114   names(ct) <- gsub("_md_assigned_only_invertebrates_", "_", substr(names(ct), 87, 168))
115   names(ct) <- gsub("_OTUs_clust", "c", names(ct))
116   names(ct) <- gsub("_tassgn", "t", names(ct))
117   names(ct) <- gsub("_tassgn", "t", names(ct))
118   names(ct) <- gsub("_AUST_", "f", names(ct))
119   names(ct) <- gsub("_SOIL_biom", "", names(ct))
120   names(ct) <- gsub("_md_assigned_only_invertebrates_", "_", names(ct))
121   return(ct)
122 }
123
124 # apply name shortening commands
125 ets_counts <- ShortenNames185(ets_counts)
126 coi_counts <- ShortenNamesCOI(coi_counts)
127
128 # function: convert count per setting to data frame suitable for plotting
129 MakeSortedDF <- function(ct){
130   ct <- data.frame(ct)
131   colnames(ct) <- "count"
132   ct[2] <- as.character(rownames(ct))
133   colnames(ct) <- c("count", "parameters")
134   # sort OTU counts per table descending
135   ct <- dplyr::arrange(ct, desc(count))
136   return(ct)
137 }
138
139 # create sorted data frames
140 ets_counts_df <- MakeSortedDF(ets_counts)
141 coi_counts_df <- MakeSortedDF(coi_counts)
142
143 # function: create plot counts versus parameters
144 PlotCountsParams <- function(cdf, data_type = c("", "18S", "COI")){
145   # melt data frame for plotting
146   melted <- melt(cdf, id.vars = "parameters")
147   # do basic barplot
148   p <- ggplot(melted, aes(x = reorder(parameters, -value), y = value)) + geom_bar(stat = "identity", fill = "grey")
149   last_plot() + theme_minimal()
150   last_plot() + theme(axis.text.y = element_text(hjust=0, angle=0), axis.text.x = element_text(hjust=0, angle=90))
151   last_plot() + ggtitle(paste("Cumulative effect of filtering parameters on", data_type, "data"))
152   last_plot() + xlab("Data sets with variable parameters")
153   last_plot() + ylab("Total phylotype count")
154 }
155
156 # plot both counts

```

```

157 plot1 <- PlotCountsParams(ets_counts_df, "18S")
158 plot2 <- PlotCountsParams(coi_counts_df, "COI")
159 grid.arrange(plot1, plot2, ncol=2)
160 }
161
162 ##### 3: get tables with highest numbers of Phylotype counts into new list
163 {
164   # finished data - lists of PhyloSeq objects:
165   psob_list_18S
166   psob_list_COI
167 }
168
169 ##### 4: Get replicate similarities during filtering
170 {
171   GetReplicateSimilarities <- function(list_item){
172     # for testing
173     # cntrl <- psob_list_18S_sel[[1]]
174
175     # get item from list
176     cntrl <- list_item
177
178     # store OTU table in matrix
179     cntrl.t <- otu_table(cntrl)
180
181     # isolate replicates via dplyr, via temporary data frame
182     soil_a <- as.matrix(dplyr::select(data.frame(cntrl.t), contains("10.H")))
183     soil_b <- as.matrix(dplyr::select(data.frame(cntrl.t), contains("11.H")))
184
185     # drop empty observations and transpose
186     soil_a <- t(soil_a[rowSums(soil_a) > 0, , drop=FALSE])
187     soil_b <- t(soil_b[rowSums(soil_b) > 0, , drop=FALSE])
188
189     # calculate distance between replicates
190     soil_a.sim <- 1-(vegdist(soil_a, method = "jaccard", na.rm = TRUE))
191     soil_b.sim <- 1-(vegdist(soil_b, method = "jaccard", na.rm = TRUE))
192     mean.sim <- mean(c(soil_a.sim, soil_b.sim))
193     sd.sim <- sd(c(soil_a.sim, soil_b.sim))
194
195     soil_ab.mean.sim <- c(soil_a.sim, soil_b.sim, mean.sim, sd.sim)
196
197     # return stress items for list items
198     return(soil_ab.mean.sim)
199   }
200
201   # apply function
202   results_a <- lapply(psob_list_18S, GetReplicateSimilarities)
203   results_b <- lapply(psob_list_COI, GetReplicateSimilarities)
204
205   results_a
206   results_b

```

Saved: 2/06/15 15:31:35

```

209
210 # apply name shortening commands
211 simis_ets <- ShortenNames18S(results_a)
212 simis_coi <- ShortenNamesCOI(results_b)
213 }
214
215 ##### 5: Combine count and similarity values
216 {
217   # function: combine results
218   # from counts and similarity comparison for plotting
219   # uses count data and results from similarity comparison
220   # generates data frame
221   CombineResults <- function(res_count, res_sim){
222     # copy input objects, in case testing is needed
223     counts.df <- res_count
224     simis.list <- res_sim
225     # covert similarity results to dataframe
226     simis.df <- t(data.frame(simis.list))
227     # add column names to dataframe
228     colnames(simis.df) <- c("soil_a", "soil_b", "mean", "sd")
229     # create combined dataframe
230     comb.df <- cbind(counts.df, simis.df)
231     # return results
232     return(comb.df)
233   }
234
235   # apply function to both data sets
236   df.csp.ets <- CombineResults(ets_counts_df, simis_ets)
237   df.csp.coi <- CombineResults(coi_counts_df, simis_coi)
238 }
239
240 df.csp.ets
241 df.csp.coi
242
243
244 ##### 6: Plot Phylotype counts and Similarity between replicates
245 {
246   # function: create plot counts versus parameters
247   PlotCPS <- function(dfcps, string = c("", "18S", "COI")){
248     #store paramters(for testing)
249     cdf <- dfcps
250     data_type <- string
251
252     # rename one dataframe column (for plotting of standard deviation) - for now removed
253     names(cdf)[names(cdf)=="sd"] <- "stdv"
254     cdf$stdv <- NULL
255
256     # melt data frame for plotting
257     melted <- melt(cdf, id.vars = "parameters")
258
259     # add variable to highlight maxima
260     melted <- mutate(melted, max = "notMax")

```

Saved: 2/06/15 15:31:35

```

261 maxs <- melted %>% group_by(variable) %>% summarise(max = max(value))
262 melted$max[which(melted$value %in% maxs$max)] <- "max"
263
264 # print
265 print(melted)
266
267 # store facet labelling alterations in lis
268 f_labs <- list(
269   'count'="Count",
270   'soil_a'="Soil 1",
271   'soil_b'="Soil 2",
272   'mean'="Soils")
273
274 # Function to rename facet labels
275 f_labeller <- function(variable,value){return(f_labs[value])}
276
277 # barplot - count data
278 p <- ggplot(melted, aes(x = reorder(parameters, -value), y = value, fill = max, group = variable))#, colour=c("red", "grey"))
279 last_plot() + facet_grid(variable~., scales = "free", labeller=f_labeller)
280 last_plot() + geom_bar(data = subset(melted, variable=="count"), stat = "identity")
281 last_plot() + geom_bar(data = subset(melted, variable=="soil_a"), stat = "identity")
282 last_plot() + geom_bar(data = subset(melted, variable=="soil_b"), stat = "identity")
283 last_plot() + geom_bar(data = subset(melted, variable=="mean"), stat = "identity")
284 # + geom_errorbar(aes(ymax = "stdv", ymin= "stdv"))
285 last_plot() + theme_bw()
286 last_plot() + theme(axis.text.y = element_text(hjust=0, angle=0), axis.text.x = element_text(hjust=0, angle=90))
287 last_plot() + ggtitle(paste("Cumulative effect of filtering parameters on", data_type, "data"))
288 last_plot() + xlab("Data sets, applied parameters")
289 last_plot() + ylab("")
290 last_plot() + guides(fill=FALSE) + scale_fill_manual(values=c("forestgreen", "grey"))
291 }
292
293 plot1 <- PlotCPS(df.csp.ets, "18S")
294 plot2 <- PlotCPS(df.csp.coi, "COI")
295 grid.arrange(plot1, plot2, ncol=2)
296 }
297 # settings with highest mean similarity between the two soils:
298 # 18S: c99 t95 f0.001
299 # COI: c97 t80 f0.001
300
301 ##### 7: Generate comparable objects of insect controls for plotting
302 {
303   # define import pathnames
304   ets_path = "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_18S_abundance/150304_0.001_discarded/150113_18_OTUs_clust97_tassgn99_md_assigned_only_no_
305   coi_path = "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_COI_abundance/150304_0.001_discarded/150121_COI_OTUs_clust97_tassgn80_md_assigned_only_ir
306
307   # Import function to get Phyloseq objects
308   # needs file list, returns biom objects in list
309   createPhyloseq <- function(path) {
310
311     # import biom file with pathname in list
312     ps_object <- try(import_biom(path))

```

```

313 # clean out strings in tax table for further handling
314
315 # remove OTUs that do not occur in any samples
316 ps_object <- prune_taxa(taxa_sums(ps_object) > 0, ps_object)
317
318 # remove crap from taxonomy strings
319 tax_table(ps_object) <- gsub("_", "", tax_table(ps_object))
320 tax_table(ps_object) <- gsub("-", "", tax_table(ps_object))
321 tax_table(ps_object) <- gsub(".", "", tax_table(ps_object))
322 tax_table(ps_object) <- gsub(" ", "", tax_table(ps_object))
323
324 # rename taxonomic ranks according to design structure
325 colnames(tax_table(ps_object))[1] = "Superphylum"
326 colnames(tax_table(ps_object))[2] = "Phylum"
327 colnames(tax_table(ps_object))[3] = "Class"
328 colnames(tax_table(ps_object))[4] = "Order"
329 colnames(tax_table(ps_object))[5] = "Family"
330 colnames(tax_table(ps_object))[6] = "Genus"
331 colnames(tax_table(ps_object))[7] = "Species"
332
333 # print taxonomy table to see what needs correction
334 print(tax_table(ps_object))
335 # return biom object
336 return(ps_object)
337 }
338
339 # read in biom files and generate a list of Phyloseq objects
340 ets <- createPhyloseq(ets_path)
341 coi <- createPhyloseq(coi_path)
342
343 ## function correct taxonomy strings in 18S and COI
344 RenameTaxa18S <- function(ets){
345   tax_table(ets)
346   tax_table(ets)[,1] <- c("Ecdysozoa")
347   tax_table(ets)[,2] <- tax_table(ets)[,4]
348   tax_table(ets)[,3] <- tax_table(ets)[,6]
349   tax_table(ets)[2,3] <- c("Eutardigrada")
350   tax_table(ets)[,4] <- c("Blattodea", NA, NA, "Lepidoptera", NA, "Hymenoptera", NA, "Odonata")
351   tax_table(ets)[,5] <- c("Blattidae", NA, NA, "Zygaenidae", NA, "Ichneumonidae", NA, "Coenagrionidae")
352   tax_table(ets)[,6] <- tax_table(ets)[,7]
353   tax_table(ets)[,7] <- tax_table(ets)[,8]
354   tax_table(ets)[,7] <- gsub(" red eyed damselfly", "", tax_table(ets)[,7])
355   tax_table(ets)[,7] <- gsub(" DJGI 2006", "", tax_table(ets)[,7])
356   tax_table(ets) <- tax_table(ets)[,-8]
357   print(tax_table(ets))
358   return(ets)
359 }
360 RenameTaxaCOI <- function(coi){
361   tax_table(coi)
362   tax_table(coi)[,1] <- tax_table(coi)[,3]
363   tax_table(coi)[,2] <- tax_table(coi)[,4]
364

```

```

365 tax_table(coi)[,3] <- tax_table(coi)[,6]
366 tax_table(coi)[,4] <- tax_table(coi)[,10]
367 tax_table(coi)[,5] <- tax_table(coi)[,11]
368 tax_table(coi)[,6] <- tax_table(coi)[,16]
369 tax_table(coi)[,7] <- tax_table(coi)[,18]
370 tax_table(coi)[,-8] -> tax_table(coi)
371 tax_table(coi)[,-8] -> tax_table(coi)
372 tax_table(coi)[,-8] -> tax_table(coi)
373 tax_table(coi)[,-8] -> tax_table(coi)
374 tax_table(coi)[,-8] -> tax_table(coi)
375 tax_table(coi)[,-8] -> tax_table(coi)
376 tax_table(coi)[,-8] -> tax_table(coi)
377 tax_table(coi)[,7] <- c(NA)
378 tax_table(coi)[10,7] <- tax_table(coi)[10,9]
379 tax_table(coi)[10,6] <- c("Acyphas")
380 tax_table(coi)[11,6] <- c("Cryptinae")
381 tax_table(coi)[6,6] <- c("Houghia")
382 tax_table(coi)[,-8] -> tax_table(coi)
383 tax_table(coi)[,-8] -> tax_table(coi)
384 tax_table(coi)[,-8] -> tax_table(coi)
385 tax_table(coi)[,-8] -> tax_table(coi)
386 tax_table(coi)[,4] <- gsub("Zygoptera", "Odonata", tax_table(coi)[,4])
387 tax_table(coi)[,4] <- gsub("Dionycha", "Araneae", tax_table(coi)[,4])
388 tax_table(coi)[,5] <- gsub("Apocrita", "Ichneumonidae", tax_table(coi)[,5])
389 tax_table(coi)[3,5] <- "Formicidae"
390 tax_table(coi)[,5] <- gsub("Glossata", "Erebidae", tax_table(coi)[,5])
391 tax_table(coi) <- gsub("NA", NA, tax_table(coi))
392
393 print(tax_table(coi))
394 return(coi)
395 }
396
397 # get clean taxonomy tables
398 ets <- RenameTaxa18S(ets)
399 coi <- RenameTaxaCOI(coi)
400
401 # function: Create object from Laurences controls:
402 newPsPb = function(physeq){
403   # Create an OTU table with one observation per species, as there are no abundance reported
404   otumat = matrix(1,nrow = 14, ncol = 1)
405   otumat
406
407   # name matrix elements
408   rownames(otumat) <- paste0("morpho", 1:nrow(otumat))
409   colnames(otumat) <- paste0("Sample", 1:ncol(otumat))
410   otumat
411
412   # generate a taxon matrix - translating only to available genetic data
413   taxmat = matrix(nrow = nrow(otumat), ncol = 7)
414
415   # name taxa according to LCs manuscript
416   taxmat[1,] <- c("Ecdysozoa", "Arthropoda", "Arachnida", "Araneae", "Sparassidae", NA, NA)

```

```

417 taxmat[2,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Blattodea", "Blattidae", "Drymaplaneta", "Drymaplaneta communis")
418 taxmat[3,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Coleoptera", "Scarabaeidae", "NA", "NA")
419 taxmat[4,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Dermoptera", "Forficulidae", "Forficula", "Forficula auricularia")
420 taxmat[5,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Diptera", "Lauxaniidae", "NA", "NA")
421 taxmat[6,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Hemiptera", "Eurybrachyidae", "NA", "NA")
422 taxmat[7,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Hymenoptera", "Formicidae", "Myrmecia", "NA")
423 taxmat[8,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Hymenoptera", "Ichneumonidae", "NA", "NA")
424 taxmat[9,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Isoptera", "Rhinoitermitidae", "Coptotermes", "NA")
425 taxmat[10,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Neuroptera", "Chrysopidae", "NA", "NA")
426 taxmat[11,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Lepidoptera", "Lymantriidae", "NA", "NA")
427 taxmat[12,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Odonata", "Coenagrionidae", "Ischnura", "Ischnura heterosticta")
428 taxmat[13,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Orthoptera", "Acrididae", "NA", "NA")
429 taxmat[14,] <- c("Ecdysozoa", "Arthropoda", "Insecta", "Orthoptera", "Tettigoniidae", "NA", "NA")
430
431 # name matrix elements
432 rownames(taxmat) <- rownames(otumat)
433 colnames(taxmat) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
434 taxmat
435
436 # convert to Phyloseq
437 OTU = otu_table(otumat, taxa_are_rows = TRUE)
438 TAX = tax_table(taxmat)
439 OTU
440
441 physeq = phyloseq(OTU, TAX)
442 physeq
443
444 #return object
445 return(physeq)
446 }
447
448 # create Phyloseq object from Laurences Paper
449 ins <- newPsb(ins)
450 }
451 tax_table(ets)
452 tax_table(coi)
453
454 otu_table(ins)
455 tax_table(ins)
456
457 ##### 7a: Plot composition of "Artificial blends"
458
459 # copy object for plotting
460 ins.p <- ins
461 otu_table(ins.p)
462 tax_table(ins.p)
463
464 # plot out different levels (Family, Genus, Species)
465 plot_bar(ins.p, "Species", fill = "Genus", facet_grid="Family~.") +
466 theme_bw() +
467 theme(axis.title.y = element_blank(), axis.text.y = element_blank(), axis.ticks.y = element_blank()) +
468 theme(strip.text.y = element_text(size=8, angle=0))

```



```
469 ##### 8: Compare similarities between insect reference and insect controls
470
471 # get info from objects for function writing
472 {
473   ets
474   coi
475   ins
476   sample_names(ets)
477   sample_names(coi)
478   sample_names(ins)
479 }
480
481 # isolate mixed sample from both samples
482 {
483   ets.mix <- prune_samples("18S.2.8.H.Insctrl",ets)
484   coi.mix <- prune_samples("COI.2.8.H.Insctrl",coi)
485 }
486
487 # remove 0 count OTUs
488 {
489   ets.mix <- prune_taxa(taxa_sums(ets.mix) > 0, ets.mix)
490   coi.mix <- prune_taxa(taxa_sums(coi.mix) > 0, coi.mix)
491 }
492
493 # rename mixed samples and reference data
494 {
495   sample_names(ets.mix) <- "18S Cntrl"
496   sample_names(coi.mix) <- "COI Cntrl"
497   sample_names(ins) <- "Ref"
498 }
499
500 # finished Phyloseq objects for testing
501 ets.mix
502 coi.mix
503 ins
504
505 sample_names(ets.mix)
506 sample_names(coi.mix)
507 sample_names(ins)
508
509 tax_table(ins)
510 tax_table(coi.mix)
511 tax_table(ets.mix)
512
513 # function: Compare reference data with otu data
514 GetMatchPerLevel <- function(refr, quer, vec_ind ){
515   # requires:
516   # - reference data as phyloseq object - subset to correct sample
517   # - query data as phyloseq object - subset to correct sample
518   # - vector index for taxonomic level
519 }
520
```

```

521 # returns:
522 # - fraction of TRUE strings contained in query when compared to reference at specific taxonomic level
523
524
525 # get input phyloseq objects for function, reference and query data
526 # - these will be function arguments
527 refr <- refr
528 quer <- quer
529 r <- vec_ind
530
531 # get names from reference OTU data, for level x
532 #r.str.lev <- na.omit(get_taxa_unique(refr, rank_names(refr)[r]))
533 #q.str.lev <- na.omit(get_taxa_unique(quer, rank_names(quer)[r]))
534
535 # get names from reference OTU data, for level x
536 r.str.lev <- get_taxa_unique(refr, rank_names(refr)[r])
537 q.str.lev <- get_taxa_unique(quer, rank_names(quer)[r])
538
539
540 print(paste("Evaluating taxonomic level: ", rank_names(refr)[r]))
541 print(paste("Reference:", sample_names(refr)))
542 print(paste("Unique taxa in reference: ", r.str.lev))
543 print(paste("Query:", sample_names(quer)))
544 print(paste("Unique taxa in query: ", q.str.lev))
545
546 # reference detected in query? - store value
547 ref.was.detected.vec <- r.str.lev %in% q.str.lev
548
549 print(paste("Query string matches to reference string: ", ref.was.detected.vec))
550
551 # calculate fraction of detected entities
552 ref.was.detected.frc <- sum(ref.was.detected.vec)/ length(ref.was.detected.vec) # length(r.str.lev)
553 print(paste("Fraction of detected entities: ", ref.was.detected.frc))
554
555 # return similarity values
556 return(ref.was.detected.frc)
557
558 }
559
560 # GetMatchPerLevel: function testing
561 GetMatchPerLevel(ins,ets,mix,1)
562 GetMatchPerLevel(ins,ets,mix,2)
563 GetMatchPerLevel(ins,ets,mix,3)
564 GetMatchPerLevel(ins,ets,mix,4)
565 GetMatchPerLevel(ins,ets,mix,5)
566 GetMatchPerLevel(ins,ets,mix,6)
567 GetMatchPerLevel(ins,ets,mix,7)
568
569 # function: GetMatchMultipleLevels
570 GetMatchMultipleLevels <- function(refr, quer){
571
572   # requires:

```

Saved: 2/06/15 15:31:35

```

573 # - reference data as phyloseq object - subset to correct sample
574 # - query data as phyloseq object - subset to correct sample
575 # - function GetMatchPerLevel()
576 # returns
577 # - vector with concordance as calculated by GetMatchPerLevel()
578
579 # get input phyloseq objects for function, reference and query data
580 # - these will be function arguments
581 refr <- refr
582 quer <- quer
583
584 # loop through rank names via vector index
585 foreach(i=1:(length(rank_names(refr))), .combine= 'c') %do% {
586   # return vector index
587   GetMatchPerLevel(refr,quer,(i))
588 }
589
590 # Compare and store in Matrix:
591 cntrl.vs.mg <- rbind(
592   GetMatchMultipleLevels(ins, ets.mix),
593   # GetMatchMultipleLevels(ets.mix, ins),
594   GetMatchMultipleLevels(ins, coi.mix)) # ,
595   # GetMatchMultipleLevels(coi.mix, ins),
596   # GetMatchMultipleLevels(coi.mix, ets.mix),
597   # GetMatchMultipleLevels(ets.mix, coi.mix))
598
599 # Name columns and rows
600 colnames(cntrl.vs.mg) <- rank_names(ins)
601 # rownames(cntrl.vs.mg) <-c("18S to Cntrl","Cntrl to COI","18S to COI", "COI to 18S")
602 rownames(cntrl.vs.mg) <-c("18S to Cntrl", "COI to Cntrl", "18S to COI", "COI to 18S")
603
604 cntrl.vs.mg
605
606
607 ##### 9: Create display Items: Insect vs Controls
608
609 ## heatmap
610 plot.new()
611 heatmap.2(cntrl.vs.mg,
612   Rowv = FALSE,
613   Colv = FALSE,
614   dendrogram = "none",
615   trace = "row",
616   tracecol = "blue",
617   # scale = "row",
618   na.rm=TRUE,
619   # sepcolor="black",
620   sepwidth=c(0.0001,0.0001),
621   colsep=1:ncol(cntrl.vs.mg),
622   rowsep=1:nrow(cntrl.vs.mg),
623   cexRow = 1.2,
624

```

```

625 cexCol = 1.4,
626 density.info=c("none"),
627 lmat = rbind(c(1,2),c(4,3)),
628 lwid = c(4,0.5),
629 lhei = c(2,.5),
630 keysize = 1.6,
631 margins = c(5,1))
632
633 ## barplot
634
635 # create data frame
636 cntrl.vs.mg.df <- data.frame(cntrl.vs.mg)
637 cntrl.vs.mg.df$Comparison <- factor(c("I8S to Cntrl.", "COI to Cntrl."))
638 cntrl.vs.mg.df
639
640 # melt data frame
641 melted <- melt(cntrl.vs.mg.df, id.vars=c("Comparison"))
642 str(melted)
643
644 # Define palette
645 cbPalette <- c( "#56B4E9", "#9999999", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC", "#E69F00" )
646
647 # create plot
648 plot3 <- ggplot(melted, aes(variable, value)) + geom_bar(aes(fill = Comparison), position = "dodge", stat="identity")
649 plot3 <- last_plot() + theme_bw()
650 plot3 <- last_plot() + xlab("") + ylab("Detected fraction of taxa contained in reference")
651 plot3 <- last_plot() + theme(axis.text=element_text(size=13))
652 plot3 <- last_plot() + theme(axis.title=element_text(size=12))
653 plot3 <- last_plot() + scale_fill_manual(values=cbPalette) + scale_colour_manual(values=cbPalette)
654 plot3
655
656 ##### 10: Create display Items: Barplot as comparsion between samples
657
658 # function: do facettted barplot
659 # works only with formatted input Phyloseq objects
660 GetBarplotsInsCntrl <- function(ins,ets,coi,mix){
661
662   # Define palette
663   cbPalette <- c( "#56B4E9", "#9999999", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC", "#E69F00" )
664
665   plot1 <- plot_bar(ins, fill="Family")
666   plot1 <- last_plot() + theme_bw()
667   plot1 <- last_plot() + facet_grid(Order~., scales = "free")
668   plot1 <- last_plot() + xlab("12 Orders, 14 Families")
669   plot1 <- last_plot() + ylab("Count")
670   plot1 <- last_plot() + theme(axis.text.y = element_text(size = 7), axis.text.x = element_blank(), axis.ticks.x = element_blank())
671   plot1
672
673   plot2 <- plot_bar(ets,mix, fill="Family")
674   plot2 <- last_plot() + theme_bw()
675   plot2 <- last_plot() + facet_grid(Order~., scales = "free")
676

```

```

677 plot2 <- last_plot() + xlab("18S: 4 Orders, 4 Families")
678 plot2 <- last_plot() + ylab("")
679 plot2 <- last_plot() + theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
680                               plot2
681
682 plot3 <- plot_bar(coi.mix, fill="Family")
683 plot3 <- last_plot() + theme_bw()
684 plot3 <- last_plot() + facet_grid(Order~., scales = "free")
685 plot3 <- last_plot() + xlab("COI: 6 Orders, 8 Families")
686 plot3 <- last_plot() + ylab("")
687 plot3 <- last_plot() + theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
688                               plot3
689
690 grid.arrange(plot1, plot2, plot3, ncol=3)
691 }
692
693 # create display items: Composition of 18S and COI Insect controls vs Reference data
694 GetBarplotsInsCntrl(ins,ets,mix,coi.mix)
695 ##### 11: Import Antarctic morphology data and metagenetic samples
696
697 # define import pathnames
698 ets_path = "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_0.001_discarded/150113_18_OTUs_clust97_tassgn99_md_assigned_only.nc"
699 coi_path = "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150304_0.001_discarded/150121_COI_OTUs_clust97_tassgn80_md_assigned_only_i
700
701 # read in biom files and generate a list of Phyloseq objects
702 ets.ant <- createPhyloseq(ets_path)
703 coi.ant <- createPhyloseq(coi_path)
704
705 # insect mix
706 ets.mix
707 coi.mix
708
709 # soil samples
710 ets.soil <- psob_list_18S[[17]]
711 coi.soil <- psob_list_COI[[8]]
712
713 # function: define new operator "not in"
714 '%!in%' <- function(x,y){ '%in%'(x,y)}
715
716 # function: filter OTUs from controls from Antarctic samples
717 # input: Phyloseq object to be filtered followed by the 2 ps object that need to be excluded
718 # returns: Filtered Phyloseq object:
719 KeepAntPS <- function(ant_ps, mix_ps, soil_ps){
720
721   # copy object
722   filtered_ps <- ant_ps
723
724   # filter samples contained in insect mix
725   otu_table(filtered_ps) <- subset(otu_table(ant_ps), subset = rownames(otu_table(mix_ps)) %!in% rownames(otu_table(mix_ps)), drop = FALSE)
726
727   # filter samples contained in soil amplicons
728

```

```

729   otu_table(filtered_ps) <- subset(otu_table(filtered_ps), subset = rownames(otu_table(filtered_ps)) %in% rownames(otu_table(soil_ps)), drop = FALSE)
730
731   # return filtered tables
732   return(filtered_ps)
733 }
734
735
736 # Filter Antarctic metagenetic samples to contain only Phylotypes not contained in any of the controls
737 ets.ant.fil <- KeepAntPS(ets.ant, ets.mix, ets.soil)
738 coi.ant.fil <- KeepAntPS(coi.ant, coi.mix, coi.soil)
739
740 ets.ant.fil
741 coi.ant.fil
742
743 # functions: correct taxonomy strings in 18S and COI
744 RenameAntTaxa18S <- function(ets.f){
745
746   # copy for function testing
747   # ets.f <- ets.ant.fil
748
749   # rename items
750   tax_table(ets.f)
751   tax_table(ets.f)[1,] <- c("Ecdysozoa", "Arthropoda", "Arachnida", "Oribatida", "Phenopelopidae", "Eupelelops", "Eupelelops plicatus", NA)
752   tax_table(ets.f)[2,] <- c("Lophotrochozoa", "Rotifera", NA, NA, NA, NA, NA)
753   tax_table(ets.f)[3,] <- c("Ecdysozoa", "Nematoda", "Chromadorea", "Araeolaimida", "Plectidae", NA, NA, NA)
754   tax_table(ets.f)[4,] <- c("Ecdysozoa", "Nematoda", "Chromadorea", "Monhysterida", "Monhysteridae", "Halomonhystera", "Halomonhystera disjuncta", NA)
755   tax_table(ets.f)[5,] <- c("Lophotrochozoa", "Rotifera", NA, NA, NA, NA, NA)
756   tax_table(ets.f) <- tax_table(ets.f)[-8]
757
758   # print and return items
759   print(tax_table(ets.f))
760   return(ets.f)
761 }
762
763 RenameAntTaxaCOI <- function(coi.f){
764
765   # rename items
766   tax_table(coi.f)
767
768   tax_table(coi.f)[1,] <- tax_table(coi.f)[3,]
769   tax_table(coi.f)[2,] <- tax_table(coi.f)[4,]
770   tax_table(coi.f)[3,] <- tax_table(coi.f)[5,]
771   tax_table(coi.f)[3,] <- gsub("Hexapoda", "Insecta", tax_table(coi.f)[3,])
772   tax_table(coi.f)[4,] <- tax_table(coi.f)[10,]
773   tax_table(coi.f)[5,] <- tax_table(coi.f)[13,]
774   tax_table(coi.f)[6,] <- tax_table(coi.f)[16,]
775   tax_table(coi.f)[7,] <- tax_table(coi.f)[17,]
776
777   tax_table(coi.f)[1,4:7] <- c("Adinetida", "Adinetidae", "Adinata", NA)
778   tax_table(coi.f)[2,4:7] <- c("Araeolaimida", "Plectidae", "Plectus", "Plectus murrayi")
779   tax_table(coi.f)[4,5:7] <- c("Geometridae", "Pleurolopha", NA)
780   tax_table(coi.f)[6,7] <- c("Pterostichus cristatus")

```

```

781 tax_table(coi.f)[9,5:7] <- c("Muscidae", "Hydrotaea", "Hydrotaea aenescens")
782 tax_table(coi.f)[10,5:7] <- c("Carabidae", "Pterostichus", "Pterostichus cristatus")
783 tax_table(coi.f)[11,5:7] <- c("Ulidiidae", "Timia", "Timia nigripes")
784 tax_table(coi.f)[12, 5:7] <- c("Sphingidae", "Cerberonoton", "Cerberonoton rubescens")
785 tax_table(coi.f)[13, 5] <- c("Geometridae")
786
787 tax_table(coi.f) <- tax_table(coi.f)[,-c(8:18)]
788
789 # print and return items
790 print(tax_table(coi.f))
791 return(coi.f)
792
793 }
794
795 # correct 18S tax strings
796 ets.ant.fil.rn <- RenameAntTaxa18S(ets.ant.fil)
797 coi.ant.fil.rn <- RenameAntTaxaCOI(coi.ant.fil)
798
799 tax_table(ets.ant.fil.rn)
800 tax_table(coi.ant.fil.rn)
801
802 # function: read in morphotype data and correct taxon and sample names
803 # works only for spiced object!
804 # return cleaned phyloseq object
805 GetAntMorph <- function(morph_path){
806
807   # import phyloseq object
808   path <- morph_path
809   morph.ant <- import_biom(path)
810
811   # create 7th column
812   tax_table(morph.ant) <- cbind(tax_table(morph.ant), NA)
813
814   # rename ranks
815   colnames(tax_table(morph.ant)) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
816   rank_names(morph.ant)
817
818   #correct OTU table
819   tax_table(morph.ant)
820   tax_table(morph.ant)[4,4:7] <- c("Parachela", "Hypsibiidae", "Acutuncus", "Acutuncus antarcticus")
821   tax_table(morph.ant)[5,4:7] <- c("Parachela", "Macrobtiidae", "Macrobtiotus", NA)
822
823   # remove OTUs that do not occur in any samples
824   morph.ant <- prune_taxa(taxa_sums(morph.ant) > 0, morph.ant)
825
826   # print object info, corrected
827   print(otu_table(morph.ant))
828   print(tax_table(morph.ant))
829
830   # return corrected phyloseq object
831   return(morph.ant)
832

```

Saved: 2/06/15 15:31:35

```

833 }
834
835 # import AVCs morpho dat and clean up
836 morph_path <- "/Users/paul/Documents/140911_c3_analysis/150302_Invertebrates/150307_AVC_OTU_table.biom"
837 morph.ant.rn <- GetAntMorph(morph_path)
838
839 # rename sample in all sets
840 sample_names(ets.ant.fil.rn) <- c("LH-1", "MS-1", "VH-2", "mix", "HI-1", "CS-2", "VH-1", "LH-2", "CS-1")
841 sample_names(coi.ant.fil.rn) <- c("HI-1", "CS-1", "LH-2", "mix", "CS-2")
842 sample_names(morph.ant.rn) <- c("CS-1", "VH-1", "LH-1", "LH-2", "HI-1", "VH-2")
843
844 sample_names(ets.ant.fil.rn)
845 sample_names(coi.ant.fil.rn)
846 sample_names(morph.ant.rn)
847
848 tax_table(morph.ant.rn)
849
850 ##### 11a: Plot composition of "Antarctic soil" reference data
851
852 # copy object for plotting
853 morph.ant.rn -> p.ant
854
855 otu_table(p.ant)
856 tax_table(p.ant)
857 sample_names(p.ant)
858
859 # plot out different levels (Family, Genus, Species)
860 plot_bar(p.ant, "Order", fill = "Class", facet_grid="Phylum~Sample") +
861   theme_bw() +
862   theme(axis.title.y = element_blank(), axis.text.y = element_blank(), axis.ticks.y = element_blank()) +
863   theme(strip.text.y = element_text(size=8, angle=0)) +
864   theme(axis.text.x = element_text(angle = 90, hjust = 1))
865
866 ##### 12: compare assignemnts
867
868 # function: subset matching samples 18S or COI vs morphotype data
869 # do pairwise comparsion
870 # requires
871 # -- three phyloseq objects with matching sample names:
872 # -- morphologic reference
873 # -- metagenetic data 1 (18S)
874 # -- metagenetic data 2 (COI)
875 # returns
876 # -- similarity matrix
877 CompareAntSamples <- function(ref, quer_a, quer_b){
878
879   # copy variables so that variable in function do not need renaming after design:
880   morph.ant.rn <- ref
881   ets.ant.fil.rn <- quer_a
882   coi.ant.fil.rn <- quer_b
883
884

```



```

885 # get samples that are available in both sets
886 ets.morph.sample.names <- sort(sample_names(morph.ant.rn)[which(sample_names(ets.ant.fil.rn))])
887 coi.morph.sample.names <- sort(sample_names(morph.ant.rn)[which(sample_names(coi.ant.fil.rn))])
888
889 # Compare 18S samples
890 # create results list
891 ResultListEts <- list()
892 # loop over vector with sample names and do comparison
893 for(i in ets.morph.sample.names){
894
895     # print(paste("Using sample: ",i))
896
897     # create new phyloseq object 1 - pruned to sample specified at vector position
898     ets.ant.fil.rn.iso <- prune_samples(i,ets.ant.fil.rn)
899     # create new phyloseq object 2 - pruned to sample specified at vector position
900     morph.ant.rn.iso <- prune_samples(i,morph.ant.rn)
901
902     # print(paste("Subset to sample: ", sample_names(ets.ant.fil.rn.iso)))
903     # print(paste("Subset to sample: ", sample_names(morph.ant.rn.iso)))
904
905     # remove OTUs that do not occur in any samples
906     ets.ant.fil.rn.iso <- prune_taxa(taxa_sums(ets.ant.fil.rn.iso) > 0, ets.ant.fil.rn.iso)
907     morph.ant.rn.iso <- prune_taxa(taxa_sums(morph.ant.rn.iso) > 0, morph.ant.rn.iso)
908
909     # print(otu_table(ets.ant.fil.rn.iso))
910     # print(otu_table(morph.ant.rn.iso))
911
912     # Compare composition
913     comp_f <- GetMatchMultipleLevels(ets.ant.fil.rn.iso, morph.ant.rn.iso)
914     # comp_r <- GetMatchMultipleLevels(morph.ant.rn.iso, ets.ant.fil.rn.iso)
915
916     # create sub-result matrix
917     s_res <- rbind(comp_f, comp_r)
918     # rownames(s_res) <- c(paste("18S", MvsG, "", i),paste("18S", GvsM, "", i))
919     # colnames(s_res) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
920
921     s_res <- rbind(comp_f)
922     rownames(s_res) <- c(paste(i, "18S"))
923     colnames(s_res) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
924
925
926     # store results in list
927     ResultListEts[[i]] <- s_res
928
929 }
930
931 # combine 18S results
932 ResultMatrixEts <- do.call(rbind, ResultListEts)
933 ResultMatrixEts
934
935 # Compare COI samples
936 # create results list
937 ResultListCoi <- list()

```

```

937 # loop over vector with sample names
938 for(i in coi.morph.sample.names){
939
940     # print(paste("Using sample: ",i))
941
942     # create new phyloseq object 1 - pruned to sample specified at vector position
943     coi.ant.fil.rn.iso <- prune_samples(i,coi.ant.fil.rn)
944     # create new phyloseq object 2 - pruned to sample specified at vector position
945     morph.ant.rn.iso <- prune_samples(i,morph.ant.rn)
946
947     # print(paste("Subset to sample: ", sample_names(coi.ant.fil.rn.iso)))
948     # print(paste("Subset to sample: ", sample_names(morph.ant.rn.iso)))
949
950     # remove OTUs that do not occur in any samples
951     coi.ant.fil.rn.iso <- prune_taxa(taxa_sums(coi.ant.fil.rn.iso) > 0, coi.ant.fil.rn.iso)
952     morph.ant.rn.iso <- prune_taxa(taxa_sums(morph.ant.rn.iso) > 0, morph.ant.rn.iso)
953
954     # print(otu_table(coi.ant.fil.rn.iso))
955     # print(otu_table(morph.ant.rn.iso))
956
957     # Compare composition
958     comp_f <- GetMatchMultipleLevels(coi.ant.fil.rn.iso, morph.ant.rn.iso)
959     # comp_r <- GetMatchMultipleLevels(morph.ant.rn.iso, coi.ant.fil.rn.iso)
960
961     # create sub-result matrix
962     s_res <- rbind(comp_f, comp_r)
963     # rownames(s_res) <- c(paste("COI", MvsG, "", i), paste("COI", GvsM, "", i))
964     # colnames(s_res) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
965
966     # create sub-result matrix
967     s_res <- rbind(comp_f)
968     rownames(s_res) <- c(paste(i, "COI"))
969     colnames(s_res) <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
970
971
972     # store results in list
973     ResultListCoi[[i]] <- s_res
974 }
975
976 # combine Coi results
977 ResultMatrixCoi <- do.call(rbind, ResultListCoi)
978 ResultMatrixCoi
979
980 # combine both results
981 ResultsComp <- rbind(ResultMatrixEts, ResultMatrixCoi)
982 ResultsComp
983
984 return(ResultsComp)
985 }
986
987 # calculate similarities at taxonomic level
988

```

```
989 res.mat <- CompareAntSamples(morph.ant.rn,ets.ant.fil.rn, coi.ant.fil.rn)
990 res.mat
991
992 # sort samples based on location
993 res.mat <- res.mat[order(substring(rownames(res.mat),1,4)),]
994
995 ## display items: generate heatmap
996 plot.new()
997 heatmap.2(res.mat,
998           Rowv = FALSE,
999           Colv = FALSE,
1000           dendrogram = "none",
1001           trace = "row",
1002           tracecol = "blue",
1003           #sepcolor="black",
1004           sepwidth=c(0.0001,0.0001),
1005           colsep=1:ncol(res.mat),
1006           rowsep=1:nrow(res.mat),
1007           cexRow = 1.4,
1008           cexCol = 1.4,
1009           density.info=c("none"),
1010           lmat = rbind(c(1,2),c(4,3)),
1011           lwid = c(4,0.5),
1012           lhei = c(2,.5),
1013           keysize = 1.6,
1014           margins = c(5,1),
1015           na.color = "grey")
1016
1017 ## display items: barplot
1018
1019 # store matrix in data frame
1020 res.mat.df <- data.frame(res.mat)
1021
1022 # name columns
1023 res.mat.df$Marker <- as.factor(substring(rownames(res.mat),6,8))
1024 res.mat.df$Location <- as.factor(substring(rownames(res.mat),1,4))
1025
1026 # filter out locations for which data is not available for both locations
1027 res.mat.df <- dplyr::filter(res.mat.df, Location != "LH-1" & Location != "VH-1" & Location != "VH-2" )
1028 str(res.mat.df)
1029
1030 # redefine factor levels
1031 res.mat.df$Location <- factor(res.mat.df$Location)
1032 res.mat.df$Marker <- factor(res.mat.df$Marker)
1033
1034 str(res.mat.df)
1035
1036 # melt data frame
1037 melted <- melt(res.mat.df, id.vars=c("Location", "Marker"))
1038 melted
1039 str(melted)
1040
```

```
1041
1042 # Define palette
1043 cbPalette <- c( "#56B4E9", "#999999", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC", "#E69F00" )
1044
1045 # barplot - count data
1046 plot4 <- ggplot(melted, aes(x = variable, y = value, fill = Marker))
1047
1048 # base plot
1049 plot4 <- last_plot() + facet_grid(Location~., scales = "free")
1050 plot4 <- last_plot() + geom_bar(data = melted ,stat = "identity", position = "dodge")
1051
1052 # make it pretty
1053 plot4 <- last_plot() + theme_bw()
1054 plot4 <- last_plot() + theme(axis.text=element_text(size=12))
1055 plot4 <- last_plot() + theme(axis.title=element_text(size=11))
1056 plot4 <- last_plot() + scale_fill_manual(values=cbPalette) + scale_colour_manual(values=cbPalette)
1057 plot4 <- last_plot() + xlab("") + ylab("Detected fraction of taxa contained in reference")
1058 plot4
1059
1060 ## do composite figure
1061
1062 # define plots
1063 upper <- plot3
1064 lower <- plot4
1065
1066 # reposition legend
1067 upper <- upper + theme(legend.justification=c(1,1), legend.position=c(1,1))
1068 lower <- lower + theme(legend.position="none")
1069
1070 # do axis titles
1071 upper <- upper + xlab("") + ylab("Matching taxonomic assignments")
1072 upper <- upper + xlab("") + ylab("Matching taxonomic assignments")
1073 theme(axis.title.x=element_text(size=12))
1074 lower <- lower + xlab("") + ylab("")
1075
1076 # re-scale x axis labels
1077 upper <- upper + theme(axis.text.x=element_text(size=13))
1078 lower <- lower + theme(axis.text.x=element_text(size=13))
1079
1080 # re-scale y axis labels
1081 upper <- upper + theme(axis.text.y=element_text(size=13))
1082 lower <- lower + theme(axis.text.y=element_text(size=9))
1083
1084 # finished plot
1085 grid.arrange(upper, lower, nrow=2)
1086
1087 ##### 13: plot taxonomic composition
1088
1089 # Barplot function
1090 GetBarplotsAnt <- function(ant,ets,coi){
1091
1092   print(paste("Omitting: ", sample_names(ets)[4]))
```

```

1093 print(paste("Omitting: ", sample_names(ets)[2]))
1094 print(paste("Omitting: ", sample_names(coi)[4]))
1095
1096 ets <- prune_samples(sample_names(ets)[-4], ets)
1097 ets <- prune_samples(sample_names(ets)[-2], ets)
1098 coi <- prune_samples(sample_names(coi)[-4], coi)
1099
1100 plot1 <- plot_bar(ant, fill="Family")
1101 plot1 <- last_plot() + theme_bw()
1102 plot1 <- last_plot() + facet_grid(Order~., scales = "free")
1103 plot1 <- last_plot() + xlab("6 Orders, 7 Families")
1104 plot1 <- last_plot() + ylab("Count")
1105 plot1 <- last_plot() + theme(axis.text.y = element_text(size = 7), axis.text.x = element_text(angle = 90, size = 6))
1106 plot1
1107
1108 plot2 <- plot_bar(ets, fill="Family")
1109 plot2 <- last_plot() + theme_bw()
1110 plot2 <- last_plot() + facet_grid(Order~., scales = "free")
1111 plot2 <- last_plot() + xlab("18S: 3 Orders, 3 Families")
1112 plot2 <- last_plot() + ylab("")
1113 plot2 <- last_plot() + theme(axis.text.y = element_text(size = 7), axis.text.x = element_text(angle = 90, size = 6))
1114 plot2
1115
1116 plot3 <- plot_bar(coi, fill="Family", facet_grid=~Order)
1117 plot3 <- last_plot() + theme_bw()
1118 plot3 <- last_plot() + facet_grid(Order~., scales = "free")
1119 plot3 <- last_plot() + xlab("COI: 5 Orders, 7 Families")
1120 plot3 <- last_plot() + ylab("")
1121 plot3 <- last_plot() + theme(axis.text.y = element_text(size = 7), axis.text.x = element_text(angle = 90, size = 8))
1122 plot3
1123
1124 grid.arrange(plot1, plot2, plot3, ncol=3)
1125 }
1126
1127 # do Barplot
1128 GetBarplotsAnt(morph.ant.rn, ets.ant.fil.rn, coi.ant.fil.rn)
1129
1130 ##### 14: Format Antarctic data into a more useful data frame
1131
1132 # store matrix in data frame
1133 res.mat.df <- data.frame(res.mat)
1134
1135 # name columns
1136 res.mat.df$Gene <- as.factor(substring(rownames(res.mat),6,8))
1137 res.mat.df$Location <- as.factor(substring(rownames(res.mat),1,4))
1138
1139 # filter out locations for which data is not available for both locations
1140 res.mat.df <- dplyr::filter(res.mat.df, Location != "LH-1" & Location != "VH-1" & Location != "VH-2" )
1141 # redefine factor levels
1142 res.mat.df$Location <- factor(res.mat.df$Location)
1143 res.mat.df
1144

```

```
1145 ##### 15: Calculate Intraclass correlation coefficient on Control and Antarctic data
1146
1147 # function: Calculate Intraclass correlation coefficient (oneway, consistency)
1148 GetICC <- function(two_df_rows){
1149
1150   test.data <- t(as.matrix(two_df_rows))
1151   print(test.data[1,1:2])
1152   print(test.data[2,1:2])
1153   print(test.data[3,1:2])
1154   print(test.data[4,1:2])
1155   print(test.data[5,1:2])
1156   print(test.data[6,1:2])
1157   print(test.data[7,1:2])
1158
1159   icc.result <- icc(test.data, model="oneway", type="consistency")
1160   print(icc.result)
1161
1162   # print(icc.result)
1163   # print(paste("value: ", icc.a$value))
1164   # print(paste("p: ", icc.a$p.value))
1165
1166   # print(paste("Genes: ", icc.a$raters))
1167   # print(paste("Ranks: ", icc.a$subjects))
1168
1169   }
1170
1171 ## Insect controls
1172 GetICC(cntrl.vs.mg) # 0.88 - 0.000976
1173
1174 ## Antarctic CS - 1
1175 GetICC(res.mat.df[1:2,1:7]) # 0.8 - 0.0166
1176
1177 ## Antarctic CS - 2
1178 GetICC(res.mat.df[3:4,1:7]) # NAN
1179
1180 ## Antarctic HI - 1
1181 GetICC(res.mat.df[5:6,1:7]) # 0.368 - 0.167
1182
1183 ## Antarctic LH - 2
1184 GetICC(res.mat.df[7:8,1:7]) # 0.853 - 0.00192
1185
1186
```

C. Analysis code chapter 4

```

1 ## 18.6.2015 - PCM invertebrates and environment - QIIME preparation
2 # =====
3 # Paul Czechowski, University of Adelaide
4 # contact: paul (dot) czechowski (at) gmail (dot) com
5 # *** code for your verification only, not for your own publications ***
6 # =====
7 # THIS SOFTWARE IS PROVIDED "AS IS" AND ANY EXPRESSED OR IMPLIED WARRANTIES,
8 # INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS
9 # FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL PAUL CZECHOWSKI (PC) OR THE UNIVERSITY OF ADELAIDE (UOFA),
10 # OR ANY OF THEIR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY,
11 # OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
12 # LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY,
13 # WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
14 # OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
15 # Without limiting the foregoing, Paul Czechowski and UOFA make no warranty that:
16 # * the software will meet your requirements.
17 # * the software will be uninterrupted, timely, secure or error-free.
18 # * the results that may be obtained from the use of the software will be effective, accurate or reliable.
19 # * the quality of the software will meet your expectations.
20 # * any errors in the software obtained web site will be corrected.
21 # Software and its documentation made available:
22 # * could include technical or other mistakes, inaccuracies or typographical errors. PC contributors may make changes
23 # * to the software or documentation made available on the web site.
24 # * may be out of date and PC, its contributors, and UOFA make no commitment to update such materials.
25 # * PC, its contributors, and USC assume no responsibility for errors or omissions in the software or documentation available.
26 # In no event shall PC, contributors, or UOFA be liable to you or any third parties for any special,
27 # punitive, incidental, indirect or consequential damages of any kind, or any damages whatsoever, including,
28 # without limitation, those resulting from loss of use, data or profits, whether or not PC,
29 # its contributors, or UOFA has been advised of the possibility of such damages, and on any theory of liability,
30 # arising out of or in connection with the use of this software.
31 # The use of the software downloaded is done at your own discretion and risk and with agreement
32 # that you will be solely responsible for any damage to your computer system or loss of data that
33 # results from such activities. No advice or information, whether oral or written, obtained by you from
34 # PC, its website, its contributors, or UOFA shall create any warranty for the software.
35 # =====
36
37 ## Steps:
38 # 1 - Get Invertebrate OTU tables that are not filtered on abundance
39 # 2 - Rename and create txt files (in case needed later)
40 # 3 - Extract sequences for biom files, align, calculate tree
41 # 4 - Collate summary counts for sequence processing

```



```

42 # 5 - Concatenate original mapping files to get observation metadata for R import
43
44
45 ## Step 1: Get Invertebrate OTU tables that are not filtered on abundance
46 #####
47 ## filter for inverts
48 ## copy to local
49 ## remove obsolete tables before further processing
50
51 ## filter for inverts and keep Phylotypes with more the 5 sequences
52 # -----
53 # 18S:
54 # ----
55 # find all Metazoans
56 target="/mnt/paul_folder/140730_18S_data/141211_18S_DeNovo/141211_18S_tax_assignment_90/141211_18S_rep_set_099_vs_db_100/141211_rep_set_0
57 awk -F '\t;' '$4 == "__Metazoa" {print $5,$6 }' "$target" | sort -d | uniq -c
58 # create output directory
59 targetdir="/mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates"
60 mkdir -p $targetdir
61 # check input file paths - choose only files with metadata and taxonomy assignments
62 ls /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/*.biom
63 # load qiime
64 module load Qiime/1.8.0
65 # loop
66 for file in /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/*.biom; do
67     [[ -e "$file" ]] || continue
68     echo @ processed file: "$file"
69     out_tmp1=$targetdir/${basename $file .biom}.tmp1
70     echo @ temp file 1 "$out_tmp1"
71     out_tmp2=$targetdir/${basename $file .biom}.tmp2
72     echo @ temp file 2 "$out_tmp2"
73     out_biom="$targetdir"/${basename "$file" .biom}_invertebrates.biom
74     echo @ biom file 2 "$out_biom"
75     echo @ retaining invertebrates
76     # filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p __Nematoda,__Rotifera,__Tardigrada,__Chelicerata
77     filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p __Nematoda,__Rotifera,__Tardigrada,__Arthropoda
78     echo @ filter empty samples
79     filter_samples_from_otu_table.py -i "$out_tmp1" -o "$out_tmp2" -n 1
80     echo @ filter 5 count OTUs
81     filter_otus_from_otu_table.py -i "$out_tmp2" -o "$out_biom" -n 5
82     echo @ generating summary files

```

```

83 biom summarize-table -i "$out_biom" -o "$targetdir"/$(basename "$out_biom" .biom).sum_qual.txt --qualitative
84 biom summarize-table -i "$out_biom" -o "$targetdir"/$(basename "$out_biom" .biom).sum_quant.txt
85 echo @ erasing temp files
86 rm -v /mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates/*.tmp?
87 done
88 # finished tables at:
89 # /mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates
90
91 # COI:
92 # ----
93 # create output directory
94 targetdir="/mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates"
95 mkdir -p $targetdir
96 # check input file paths - choose only files with metadata and taxonomy assignments
97 ls /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/*md_assigned_only.biom
98 # load qiime
99 module load Qiime/1.8.0
100 # loop
101 for file in /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/*md_assigned_only.biom; do
102     [[ -e "$file" ]] || continue
103     echo @ processed file: "$file"
104     out_tmp1=$targetdir/$(basename $file .biom).tmp1
105     echo @ temp file 1 "$out_tmp1"
106     out_tmp2=$targetdir/$(basename $file .biom).tmp2
107     echo @ temp file 2 "$out_tmp2"
108     out_biom="$targetdir"/$(basename "$file" .biom)_invertebrates.biom
109     echo @ biom file 2 "$out_biom"
110     echo @ retaining invertebrates
111     # filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p Chelicerata,Nematoda,Rotifera,Tardigrada
112     filter_taxa_from_otu_table.py -i "$file" -o "$out_tmp1" -p Arthropoda,Nematoda,Rotifera,Tardigrada
113     echo @ filter empty samples
114     filter_samples_from_otu_table.py -i "$out_tmp1" -o "$out_tmp2" -n 1
115     echo @ filter 5 count OTUs
116     filter_otus_from_otu_table.py -i "$out_tmp2" -o "$out_biom" -n 5
117     echo @ generating summary files # update pathnames
118     biom summarize-table -i "$out_biom" -o "$targetdir"/$(basename "$out_biom" .biom).sum_qual.txt --qualitative
119     biom summarize-table -i "$out_biom" -o "$targetdir"/$(basename "$out_biom" .biom).sum_quant.txt
120     echo @ erasing temp files
121     rm -v /mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates/*.tmp?
122 done
123 # finished tables at:

```

```

124 # /mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates
125
126 ## copy to local
127 # -----
128 # 18S:
129 local_file="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_from_starcluster/150507_18S"
130 remote_file="/mnt/paul_folder/140730_18S_data/150302_18S_phylotypes_invertebrates/"
131 scp -r -C -i "$remote_file" "$local_file"
132 # COI:
133 local_file="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_from_starcluster/150507_COI"
134 remote_file="/mnt/paul_folder/140719_COI_data/150302_COI_phylotypes_invertebrates/"
135 scp -r -C -i "$remote_file" "$local_file"
136
137 ## remove obsolete tables before further processing
138 # -----
139 # these use the relatively strict settings from Chapter 3
140 # delete obsolete 18S tables (not 97-70)
141 find /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150511_from_starcluster/150511_18S -type f ! -iname "150113_18_OTUs_clus
142 # delete obsolete COI tables (not 97-75)
143 find /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150511_from_starcluster/150511_COI -type f ! -iname "150121_COI_OTUs_tlu
144
145 ## Step 2: rename and create txt files (in case needed later)
146 ## =====
147 # filtering is not used in this step, as it would remove
148 # a lot of Antarctic Phylotypes with the controls left in
149
150 # input file lists
151 table[1]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150511_from_starcluster/150511_18S/150113_18_OTUs_clust97_tassgn97_n
152 table[2]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150511_from_starcluster/150511_COI/150121_COI_OTUs_clust97_tassgn75_
153
154 # mapping files
155 mf[1]="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_18S_MF.txt"
156 mf[2]="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_COI_MF.txt"
157
158 # define path and filenames
159 outdir[1]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_18S"
160 outdir[2]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_COI"
161
162 # loop over tables
163 for ((j = 1; j <= "${#table[@]}"; j++)); do
164     echo "1 --- using table: ${table[j]}"

```

```

165 # filename definition, abundance filtering
166 filtered="{outdir[j]}/150511_UF_${basename "${table[j]}"}.biom).tmp
167 echo "2 -- $filtered"
168 filter_otus_from_otu_table.py -i "${table[j]}" -o "$filtered" -s 1
169 # filename definition, remove empty samples
170 filtered_s="{outdir[j]}/150511_UF_${basename "${table[j]}"}.biom).biom
171 echo "3 -- $filtered_s"
172 mf_out="{outdir[j]}/150511_UF_${basename "${mf[j]}"}.txt).txt
173 echo "4 -- $mf_out"
174 filter_samples_from_otu_table.py -i "$filtered" -o "$filtered_s" -m "${mf[j]}" --output_mapping_fp "${mf_out}" -n 1
175 # filename definition and summary
176 sumfile=$(dirname $filtered_s)/${basename $filtered_s}.biom)
177 echo "5 -- $sumfile"
178 biom summarize-table -i "$filtered_s" -o "$sumfile".sum_qual.txt --qualitative
179 biom summarize-table -i "$filtered_s" -o "$sumfile".sum_quant.txt
180 # filename definition and convert to text
181 filtered_s_txt=$(dirname "$filtered_s")/${basename "$filtered_s".biom).txt
182 echo "6 -- $filtered_s_txt"
183 biom convert -i "$filtered_s" -o "$filtered_s_txt" -b --header-key taxonomy --output-metadata-id "Taxonomy"
184 done
185
186 ## Step 2a: Look at taxonomy strings
187 ## =====
188
189 # input txt files
190 ls /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_18S/150511_UF_150113_18_OTUs_clust97_taxgr
191 ls /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_COI/150511_UF_150121_COI_OTUs_clust97_taxgr
192
193 # Excel file
194 ls /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_tax_strings.xlsx
195
196
197 ## Step 2b: create *copies* of .biom for better handling
198 ## =====
199 # these are for R import
200 ls /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_18S/150511_18S.biom
201 ls /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_COI/150511_COI.biom
202
203
204 ## Step 3: Extract sequences for biom files, align, calculate tree
205 ## =====

```

```
206
207 # get repset 18S to local
208 local_file="/Users/paul/Documents/150127_c4_analysis/150406_18S_c97_repset.fasta" # use full paths with leading slash
209 remote_file="/mnt/paul_folder/140730_18S_data/150112_18S_OTU_rep_sets/150113_rep_set_097.fasta" # use full paths with leading slash
210 scp -i "$remote_file" "$local_file"
211
212 # get repset COI to local
213 local_file="/Users/paul/Documents/150127_c4_analysis/150403_repsets/150406_COI_c97_repset.fasta" # use full paths with leading slash
214 remote_file="/mnt/paul_folder/140719_COI_data/140925_rep_set/140825_rep_set_097.fasta" # use full paths with leading slash
215 scp -i "$remote_file" "$local_file"
216
217 # define input tables
218 table[1]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_18S/150511_18S.biom"
219 table[2]="/Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150511_COI/150511_COI.biom"
220
221 # repsets - checked
222 repset[1]="/Users/paul/Documents/150127_c4_analysis/150403_repsets/150406_18S_c97_repset.fasta"
223 repset[2]="/Users/paul/Documents/150127_c4_analysis/150403_repsets/150406_COI_c97_repset.fasta"
224
225 # sequence sets - output files
226 seq[1]="/Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_18S_seqs.fasta"
227 seq[2]="/Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_COI_seqs.fasta"
228
229 # sequence sets - output files - touching only - fill with Geneious
230 algn[1]="/Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_18S_seqs_algn.fasta"
231 algn[2]="/Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_COI_seqs_algn.fasta"
232
233 # extract sequence sets for alignment
234 for ((i = 1; i <= "${#table[@]}"; i++)); do
235     [[ -e ${table[i]} ]] || continue
236     filter_fasta.py -b "${table[i]}" -f "${repset[i]}" -o "${seq[i]}"
237     touch "${algn[i]}"
238 done
239
240
241 # unaligned sequence files
242 # -----
243 # 18S:
244 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_18S_seqs.fasta
245 # COI:
246 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_COI_seqs.fasta
```

```
247 # unaligned sequence files - with shorter names for R import
248 # -----
249 # 18S:
250 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_18S_seqs_sn.fasta
251 # COI:
252 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_COI_seqs_sn.fasta
253 # alignment and tree files:
254 # -----
255 ## alignments done in Geneious 7.2.1 - folder 150507
256 # MAFFT 7.017 - auto - preserve order, determine direction, adjust direction
257 # 18S:
258 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_18S_seqs_algn.fasta
259 # COI:
260 ls /Users/paul/Documents/150127_c4_analysis/150403_sequence_sets/150511_COI_seqs_algn.fasta
261 # trees - done in Geneious 7.2.1 with FastTree 2.1.5 - GTR
262 # 18S:
263 ls /Users/paul/Documents/150127_c4_analysis/150403_trees/150511_18S_fasttree.tre
264 # COI:
265 ls /Users/paul/Documents/150127_c4_analysis/150403_trees/150511_COI_fasttree.tre
266 # Step 4: collate summary counts
267 # =====
268 # all phylotypes
269 # -----
270 # 18S
271 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables/150113_18_OTUs_clust97_tassgn90_md.sum_qual.txt
272 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables/150113_18_OTUs_clust97_tassgn90_md.sum_quant.txt
273 # COI
274 cat /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/150121_COI_OTUs_clust97_tassgn75_md.sum_qual.txt
275 cat /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/150121_COI_OTUs_clust97_tassgn75_md.sum_quant.txt
276 # assigned only
277 # -----
```

```
288
289 # 18S
290 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables/150113_18_OTUs_clust97_tassgn90_md_assigned_only.sum_qual.txt
291 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables/150113_18_OTUs_clust97_tassgn90_md_assigned_only.sum_quant.txt
292
293 # COI
294 cat /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/150121_COI_OTUs_clust97_tassgn75_md_assigned_only.sum_qual.txt
295 cat /mnt/paul_folder/140719_COI_data/151021_COI_OTU_tables/150121_COI_OTUs_clust97_tassgn75_md_assigned_only.sum_quant.txt
296 # after blank subtraction - 18S only
297 # -----
298
299 # 18S, Starcluster
300 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/150113_18_OTUs_clust97_tassgn90_md_assigned_only_no_contain
301 cat /mnt/paul_folder/140730_18S_data/150113_18S_OTU_tables_subtracted_blanks/150113_18_OTUs_clust97_tassgn90_md_assigned_only_no_contain
302
303 # COI, no data
304
305 # sub-setting to invertebrates and 5+ seqs
306 # -----
307
308 # 18S
309 cat /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150507_18S/150507_UF_150113_18_OTUs_clust97_tas
310 cat /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150507_18S/150507_UF_150113_18_OTUs_clust97_tas
311
312 # COI
313 cat /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150507_COI/150507_UF_150121_COI_OTUs_clust97_tas
314 cat /Users/paul/Documents/150127_c4_analysis/150401_OTU_tables/150507_unfiltered_subset/150507_COI/150507_UF_150121_COI_OTUs_clust97_tas
315
316 # Step 5 - Concatenate original mapping files to get observation metadata for R import
317 # =====
318
319 # mapping files
320 mf_a="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_18S_MF.txt"
321 mf_b="/Users/paul/Documents/140911_c3_analysis/mapping_files/150107_mf_metadata/150108_COI_MF.txt"
322
323 cat "$mf_a" "$mf_b" > /Users/paul/Documents/150127_c4_analysis/150512_observation_metadata/150512_observation_md.txt
```

```
1 ## 18.6.2015 - PCM invertebrates and environment - R analysis
2 # =====
3 # Paul Czechowski, University of Adelaide
4 # contact: paul (dot) czechowski (at) gmail (dot) com
5 # *** code for your verification only, not for your own publications ***
6 # =====
7 # THIS SOFTWARE IS PROVIDED "AS IS" AND ANY EXPRESSED OR IMPLIED WARRANTIES,
8 # INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS
9 # FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT PAUL CZECHOWSKI (PC) OR THE UNIVERSITY OF ADELAIDE (UOFA),
10 # OR ANY OF THEIR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY,
11 # OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
12 # LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY,
13 # WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
14 # OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
15 # Without limiting the foregoing, Paul Czechowski and UOFA make no warranty that:
16 # * the software will meet your requirements.
17 # * the software will be uninterrupted, timely, secure or error-free.
18 # * the results that may be obtained from the use of the software will be effective, accurate or reliable.
19 # * the quality of the software will meet your expectations.
20 # * any errors in the software obtained web site will be corrected.
21 # Software and its documentation made available:
22 # * could include technical or other mistakes, inaccuracies or typographical errors. PC contributors may make changes
23 # * to the software or documentation made available on the web site.
24 # * may be out of date and PC, its contributors, and UOFA make no commitment to update such materials.
25 # * PC, its contributors, and USC assume no responsibility for errors or omissions in the software or documentation available.
26 # In no event shall PC, contributors, or UOFA be liable to you or any third parties for any special,
27 # punitive, incidental, indirect or consequential damages of any kind, or any damages whatsoever, including,
28 # without limitation, those resulting from loss of use, data or profits, whether or not PC,
29 # its contributors, or UOFA has been advised of the possibility of such damages, and on any theory of liability,
30 # arising out of or in connection with the use of this software.
31 # The use of the software downloaded is done at your own discretion and risk and with agreement
32 # that you will be solely responsible for any damage to your computer system or loss of data that
33 # results from such activities. No advice or information, whether oral or written, obtained by you from
34 # PC, its website, its contributors, or UOFA shall create any warranty for the software.
35 # =====
36
37 ##### Steps: functions are defined in Step 5 or in the beginning of Steps 1,2,3,4):
38 # 0 - Environment setup
39 # 1 - Data import
40 # a - Phyloseq objects
41 # b - Observation metadata
```



```
42 # 2 - Determination of suitable predictor variables and subsetting and plotting
43 # a - initial PCA to show variables
44 # b - identify and remove cocolated variables and outliers
45 # c - another PCA to show data after cleaning - here also evaluation of environmental PCAs for manuscript (and calc of new PCs with care
46 # d - create metadata object that will be used for observation metadata
47 #
48 # 3 - Filtering and cleaning of phylotype data sets
49 # a - subset datasets to PCM data only, substract Australian controls, remove Antarctic controls
50 # b - Gene-specific cleanup of taxonomy tables - remove insects - correct taxonomy - remove Arthropods - remove RH region
51 # c - Remove empty samples to increase data density
52 # d - Check finished unmerged objects
53 #
54 # 4 - merge Phylotype data and sample data into one objects
55 # a - prepare merging
56 # b - merge OTU and TAX tables
57 # c - merge *subsetting / preprocessed* observation metadata into sample data
58 # d - create Phyloseq object and collect garbage
59 #
60 # 5 - Functions for Analysis
61 # a - Simple maps for testing
62 # b - Simple NMDS for testing
63 # c - Taxon agglomeration - know what you put in here first!
64 # e - Subset to Taxonomic Level
65 # f - Wrapper function to convert OTU tables for Vegan
66 # g - Calculate distance matrices - return list
67 # h - calculate MDS of biotic data
68 # i - Get Observation and Species data
69 # k - Return Models 1 and 0
70 # l - Get Models with best AIC score for Constrained ordinations
71 # m - Plot CA items in lists, needs Models in List and List with Phyloseq objects and some other things
72 # n - Get List with subsetting to taxa pairs
73 # o - Plot MDS - could be combined with Plot CA with little work
74 #
75 # 6 - Constrained ordinations: CCA / RDA / Capscale
76 # a - Tax glom at class level
77 # b - plot Phyloseq object and also do a PCA (to see what looks better)
78 # c - Get list of Phyloseq objects, for each Phylum level
79 # d - Get Species occurrence and Observation metadata
80 # e - Calculate empty and full models
81 # f - Find highest scoring Models for all Constrained Ordination types
```

Saved: 18/06/15 16:11:14

```
83 # g - plot objects
84 # h - evaluate model
85 # i - generate counts for results section
86
87 # 8 - data export
88 # a - data export of coordinates for Duanne in KML format
89 # b - data export of sample coordinates for main text figure
90
91 # 9 - counts for writing
92 # a - total sample number
93 # b - biologic data per Gene
94
95
96
97 ##### 0 - Environment setup
98 ## clear environment, set working directory
99 rm(list=ls()) # clear R environment
100 setwd("")
101
102 ## load packages
103 require("phyloseq")
104 require("biom")
105 require("plyr")
106 require("dplyr")
107 require("vegan")
108 require("gplots")
109 require("ade4")
110 require("ggplot2")
111 require("reshape2")
112 require("gridExtra")
113 require("foreach")
114 require("caret")
115 require("corrplot")
116 require("ggbiplot")
117 require("outliers")
118 require("caret")
119 require("phylogeo")
120 require("xlsx")
121 require("sp")
122 require("rgdal")
123
```

```

124 ## get package citations
125 cap_cit <- FALSE
126 if(cap_cit == TRUE){
127   capture.output(utils::print.bibentry(citation("phylloseq"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
128   capture.output(utils::print.bibentry(citation("biom"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
129   capture.output(utils::print.bibentry(citation("biom"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
130   capture.output(utils::print.bibentry(citation("plyr"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
131   capture.output(utils::print.bibentry(citation("dplyr"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
132   capture.output(utils::print.bibentry(citation("vegan"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
133   capture.output(utils::print.bibentry(citation("gplots"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
134   capture.output(utils::print.bibentry(citation("ade4"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
135   capture.output(utils::print.bibentry(citation("ggplot2"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
136   capture.output(utils::print.bibentry(citation("reshape2"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
137   capture.output(utils::print.bibentry(citation("gridExtra"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
138   capture.output(utils::print.bibentry(citation("foreach"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
139   capture.output(utils::print.bibentry(citation("caret"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
140   capture.output(utils::print.bibentry(citation("corrplot"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
141   capture.output(utils::print.bibentry(citation("ggbiplot"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
142   capture.output(utils::print.bibentry(citation("outliers"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
143   capture.output(utils::print.bibentry(citation("caret"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
144   capture.output(utils::print.bibentry(citation("phylogeo"), style = "Bibtex"), file = "150603_C4_package_list.bib", append = TRUE)
145 }
146
147
148 ##### 1 - Data import (biom, sequences, trees, observation metadata)
149
150 ## a - Phyloseq objects
151
152 # 18S file paths:
153 path.e.b = "/150511_18S.biom"
154 path.e.t = "/150511_18S_fasttree.tre"
155 path.e.s = "150511_18S_seqs_sn.fasta"
156
157 # COI file paths:
158 path.c.b = "/150511_COI.biom"
159 path.c.t = "/150511_COI_fasttree.tre"
160 path.c.s = "150511_COI_seqs_sn.fasta"
161
162 # Import
163 ps.e <- import_biom(path.e.b, treefilename = path.e.t, refseqfilename = path.e.s)
164 ps.c <- import_biom(path.c.b, treefilename = path.c.t, refseqfilename = path.c.s)

```

Saved: 18/06/15 16:11:14

```

165
166 # needed for debugging a lot later
167 ps.e.original <- ps.e
168 ps.c.original <- ps.c
169
170 # garbage collection
171 rm(path.e.b, path.e.t, path.e.s, path.c.b, path.c.t, path.c.s)
172
173 ## b - Observation metadata for determination of predictor variables and later merged Phyloseq object
174
175 # read in observation data
176 # - remove duplicates - name rows - check amounts of samples
177 # - remove samples from Reinbolt Hills
178
179 md <- read.csv("/150521_observation_md.csv", stringsAsFactors = FALSE)
180 md <- md[!duplicated(md$Sample),]
181 rownames(md) <- md$Sample
182
183 # remove Reinbolt Hills
184 md -> md.original
185 # md <- md.original
186 md <- md[(md$Location != "Reinbolt_Hills"),]
187
188 # xrd data read in - scale to sum 1 by row
189 md[grep("x_", colnames(md))] <- t(apply(md[grep("x_", colnames(md))], 1, function(x) {x/sum(x)}))
190
191 # check out labeling - store sample names and variable labels of metadata for PCA biplots
192 x.sample.names <- rownames(na.omit(md[grep("x_", colnames(md))])) # xrd sample names
193 c.sample.names <- rownames(na.omit(md[grep("c_", colnames(md))])) # csbp sample names
194
195 # export as xlsx to wd for supplemental materials - see http://www.r-bloggers.com/importexport-data-to-and-from-xlsx-files/
196 # getwd()
197 # write.xlsx(md, "150615_observation_metadata.xlsx")
198
199 ##### 2 - Determination of suitable predictor variables and subsetting
200
201 # function - remove outliers and replace with mean - (Komsta, L. (2006). Processing data for outliers. R News, 6(2), 10-13.)
202 rmOut <- function(x) {
203   x <- as.matrix(x)
204   print(outlier(x, logical=TRUE))
205   x <- rm.outlier(x, fill = TRUE)

```

```
206     return(data.frame(x))
207 }
208
209 # function - preprocessing with carret package - set up preprocessing variables
210 Trans <- function(x) {
211     # get transformation parameters - no centering
212     x.pp <- preProcess(x, method=c("BoxCox", "scale", "center"), verbose = TRUE)
213     # x.pp <- preProcess(x, method=c("range"), verbose = TRUE)
214     print(x.pp)
215     # do transformation of objects
216     y <- predict(x.pp, x)
217     # return function values
218     return(y)
219 }
220
221 # function: check for cocrrelation among xrd variables
222 RmCor <- function(x, ptile){
223
224     # calculate initial correlation
225     cor(x) -> cor.x
226
227     # plot initial correlations
228     corplot(cor.x, type = "lower", order = "AOE", cl.pos = "b", tl.pos = "d", tl.col = "black", tl.srt = 60)
229     title(c(ptile), line = -2)
230
231     # find highly correlated values
232     hc.x <- findCorrelation(cor.x, cutoff=0.75)
233
234     # if highly correlated values are there
235     if(length(c(hc.x))!= 0){
236
237         # print highly correlated
238         print(hc.x)
239         print(colnames(x[,c(hc.x)]))
240
241         # remove highly correlated
242         x.nc <- x[,~c(hc.x)]
243
244         # calculate new correlations
245         cor(x.nc) -> cor.x.nc
246     }
```

```

247 # plot new correlation
248 corplot(cor.x.nc, type = "lower", order = "AOE", cl.pos = "b", tl.pos = "d", tl.cex = 0.7, tl.col = "black", tl.srt = 60,
249 # title(c(ptitle), line = -2)
250 }
251
252 else
253 {
254 # just use the original data
255 x.nc <- x
256 }
257
258 # return finished object
259 return(x.nc)
260 }
261
262 ## a - PCA on correlation matrix (scales / centered) to detect cocrrelated variables
263
264 # create input data - remove Reinbolt Hills
265
266 pca.x.input <- na.omit(md[grep("x_", colnames(md))]) # xrd data
267 pca.c.input <- na.omit(md[grep("c_", colnames(md))]) # csbp data
268
269 # adding more things - complete dataset, saves a lot of time later:
270 names(md)
271
272 # define variables that are to be used for preprocessing - see metadata read in
273 var.names <- c("c_Gravel", "c_Texture", "c_Ammonium", "c_Nitrate", "c_Phosphorus",
274 "c_Potassium", "c_Sulphur", "c_Org_Carbon", "c_Conductivity", "c_pH_CaCl2",
275 "c_pH_H2O", "x_Quartz", "x_Feltspar", "x_Titanite", "x_Pyr_Amp_Gar", "x_Micas",
276 "x_Dolomite", "x_Kao_Chlor", "x_Calcite", "x_Chlorite", "g_Elevation",
277 "s_Slope", "t_Soil_Temp")
278
279 # select sample data column indices that are to be used for preprocessing
280 var.ind.touched <- which(names(md) %in% var.names)
281
282 # select sample data column indices that are to be used for preprocessing
283 pca.xc.input <- na.omit(md[,var.ind.touched]) # all data deemed important
284
285 # correct variable labels
286 colnames(pca.x.input) <- substr(colnames(pca.x.input), 3, 15) # xrd - good variable labels
287 colnames(pca.c.input) <- substr(colnames(pca.c.input), 3, 15) # csbp - good variable labels

```

```

288 colnames(pca.xc.input) <- substr(colnames(pca.xc.input),3, 15) # csbp - good variable lables
289
290 # export combined data table for supplemental information
291 # setwd("~/Documents/15_Ph_D_thesis_C4_Invertebrates_Environment/150610_ms_dev/150603_display_items/150610_raw/150615_tables")
292 # write.xlsx(pca.xc.input, "150615_raw_md.xlsx")
293 # sink("150615_raw_md_sum.txt", type = "output")
294 # summary(pca.xc.input)
295 # sink()
296 # setwd("")
297
298 # get sampling locations as factors - for biplots
299 x.loc.group <- as.factor(md[which(rownames(pca.x.input)%in% rownames(md)),2])
300 c.loc.group <- as.factor(md[which(rownames(pca.c.input)%in% rownames(md)),2])
301 xc.loc.group <- as.factor(md[which(rownames(pca.xc.input)%in%rownames(md)),2])
302
303 # calculate PCAs on correlation matrices
304 pca.x <- prcomp(pca.x.input, center = TRUE, scale = TRUE) # xrd data
305 pca.c <- prcomp(pca.c.input, center = TRUE, scale = TRUE) # csbp data
306 pca.xc <- prcomp(pca.xc.input, center = TRUE, scale = TRUE) # csbp data
307
308 # create biplots
309 pca.xc.p <- ggbiplot(pca.xc,
310                     choices = 1:2,
311                     scale = 1,
312                     pc.biplot = TRUE,
313                     ellipse = TRUE,
314                     varname.size = 3,
315                     groups = xc.loc.group,
316                     # labels = c.sample.names,
317                     ) + ggtitle("A: All variables") + theme_bw() + theme(plot.title = element_text(hjust=0), legend.title=element_b
318
319 pca.x.p <- ggbiplot(pca.x,
320                     choices = 1:2,
321                     scale = 1, pc.biplot = TRUE,
322                     ellipse = TRUE,
323                     varname.size = 3,
324                     groups = x.loc.group,
325                     # labels = x.sample.names,
326                     ) + ggtitle("B: Minerals") + theme_bw() + coord_flip() + theme(plot.title = element_text(hjust=0), legend.positi
327
328 pca.c.p <- ggbiplot(pca.c,
329                     choices = 1:2,
330                     scale = 1,

```

```
329     pc.biplot = TRUE,
330     ellipse = TRUE,
331     varname.size = 3,
332     groups = c.loc.group,
333     # labels = c.sample.names,
334     ) + ggtitle("C: Geochemistry") + theme_bw() + theme(plot.title = element_text(hjust=0), legend.position="none")
335 grid.arrange(pca.xc.p, pca.x.p, pca.c.p, ncol = 3)
336
337 ## b - identify and remove outliers and cocorrelated variables
338
339 # using data from initial PCA analysis
340 fil.x <- pca.x.input
341 fil.c <- pca.c.input
342 fil.xc <- pca.xc.input
343
344 # remove outliers
345 fil.x.no <- rmOut(fil.x)
346 fil.c.no <- rmOut(fil.c)
347 fil.xc.no <- rmOut(fil.xc)
348
349 # apply preprocessing function
350 fil.x.no.tr <- Trans(fil.x.no)
351 fil.c.no.tr <- Trans(fil.c.no)
352 fil.xc.no.tr <- Trans(fil.xc.no)
353
354 # remove cocorrelated
355 plot.new()
356 par(mfrow = c(3,2))
357 fil.xc.no.tr.ss <- RmCor(fil.xc.no.tr, "All variables")
358 fil.c.no.tr.ss <- RmCor(fil.c.no.tr, "Chemistry")
359 fil.x.no.tr.ss <- RmCor(fil.x.no.tr, "Minerals")
360
361 # for manuscript
362 par(mfrow = c(1,2))
363 fil.xc.no.tr.ss <- RmCor(fil.xc.no.tr, "All variables")
364
365 # export combined data table for supplemental information
366 # setwd("")
367 # write.xlsx(fil.xc.no.tr.ss, "150615_pp_md.xlsx")
368 # sink("150615_pp_md_sum.txt", type = "output")
369 # summary(fil.xc.no.tr.ss)
```



```

370 # sink()
371 # setwd("")
372
373 ## c - another PCA to show data after cleaning
374
375 # get sampling locations as factors
376 x.ss.loc.group <- as.factor(md[which(rownames(fil.x.no.tr.ss)%in% rownames(md)),2])
377 c.ss.loc.group <- as.factor(md[which(rownames(fil.c.no.tr.ss)%in% rownames(md)),2])
378 xc.ss.loc.group <- as.factor(md[which(rownames(fil.xc.no.tr.ss)%in% rownames(md)),2])
379
380
381 # calculate PCAs on correlation matrices
382 pca.x.ss <- prcomp(fil.x.no.tr.ss, center = TRUE, scale = TRUE) # xrd data
383 pca.c.ss <- prcomp(fil.c.no.tr.ss, center = TRUE, scale = TRUE) # csbp data
384 pca.xc.ss <- prcomp(fil.xc.no.tr.ss, center = TRUE, scale = TRUE) # csbp data
385
386 # looking at PCAs before plotting, for manuscript
387
388 # get variance plots for PCAs
389 par(mfrow = c(1,1))
390 plot.new()
391 plot(pca.x.ss, type = "l")
392 plot(pca.c.ss, type = "l")
393 plot(pca.xc.ss, type = "l")
394
395 # inspect original PCA objects (transformed, scaled cntred, but with all original PCs)
396 pca.x.ss$rotation
397 pca.c.ss$rotation
398 pca.xc.ss$rotation
399
400 # check information content using caret package - retains PCs with 95% variation content
401 # see objects
402 trans_x <- preProcess(fil.x.no.tr.ss, method = "pca")
403 trans_c <- preProcess(fil.c.no.tr.ss, method = "pca")
404 trans_xc <- preProcess(fil.xc.no.tr.ss, method = "pca")
405
406 # check loadings of new PCs - looks the same as the old ones except last PCs were discarded
407 trans_x$rotation
408 trans_c$rotation
409 trans_xc$rotation
410

```

```
411 # calculate new PCAs (not used in manuscript)
412 PC1 <- predict(trans_x,fil.x.no.tr.ss)
413 PC2 <- predict(trans_c,fil.c.no.tr.ss)
414 PC3 <- predict(trans_xc,fil.xc.no.tr.ss)
415
416
417 # create biplots
418 pca.xc.p.ss <- ggbiplot(pca.xc.ss,
419   choices = 1:2,
420   scale = 1,
421   pc.biplot = TRUE,
422   ellipse = TRUE,
423   varname.size = 3,
424   groups = xc.ss.loc.group,
425   # labels = c.sample.names,
426   ) + ggtitle("All variables") + theme_bw() + theme(legend.title=element_blank(), legend.background = element_rect
427   choices = 1:2,
428   scale = 1, pc.biplot = TRUE,
429   ellipse = TRUE,
430   varname.size = 3,
431   groups = x.ss.loc.group,
432   # labels = x.sample.names,
433   ) + ggtitle("Minerals") + theme_bw() + theme(legend.position="none")
434
435 pca.c.p.ss <- ggbiplot(pca.c.ss,
436   choices = 1:2,
437   scale = 1,
438   pc.biplot = TRUE,
439   ellipse = TRUE,
440   varname.size = 3,
441   groups = c.ss.loc.group,
442   # labels = c.sample.names,
443   ) + ggtitle("Chemistry") + theme_bw() + theme(legend.position="none")
444
445 grid.arrange(pca.xc.p.ss, pca.x.p.ss, pca.c.p.ss, ncol = 3)
446
447 # again, for manuscript, single plot;
448 plot.new()
449 env_pca <- ggbiplot(pca.xc.ss,
450   choices = 1:2,
451   scale = 1,
452   pc.biplot = TRUE,
```

```

452 ellipse = TRUE,
453 varname.size = 4,
454 groups = xc.ss.loc.group,
455 # labels = c.sample.names,
456 ) + theme_bw() + theme(legend.title=element_blank(), legend.background = "gray90",
457 # and double plot for supplement
458 pca.x.p.ss <- ggbiplot(pca.x.ss,
459 choices = 1:2,
460 scale = 1, pc.biplot = TRUE,
461 ellipse = TRUE,
462 varname.size = 4,
463 groups = x.ss.loc.group,
464 # labels = x.sample.names,
465 ) + ggtitle("A: Minerals") + theme_bw() + theme(plot.title = element_text(hjust=0), legend.position="hor
466
467 pca.c.p.ss <- ggbiplot(pca.c.ss,
468 choices = 1:2,
469 scale = 1,
470 pc.biplot = TRUE,
471 ellipse = TRUE,
472 varname.size = 4,
473 groups = c.ss.loc.group,
474 # labels = c.sample.names,
475 ) + ggtitle("B: Geochemistry") + theme_bw() + theme(plot.title = element_text(hjust=0), legend.title=ele
476
477 grid.arrange(pca.x.p.ss, pca.c.p.ss, ncol = 2)
478
479 ## d - create metadata object that will be used for observation metadata
480
481 # using filtered, scaled, centered xrd and data from previous steps
482 md.fsc <- data.frame(fil.xc.no.tr.ss)
483
484 # create modified version of original metadata: select columns in the original data that are contained in the filtered data, and columns
485 md.o.ss <- select(data.frame(md[c(which(rownames(md) %in% rownames(md.fsc)))]), one_of(c("Sample", "Location", "Genes", "CSBP_id", "XRF
486
487 # merge dataframes by row names, assign rownames again, assign sample names
488 md.fsc.complete <- merge(md.fsc, md.o.ss, by=0, all=TRUE )
489 rownames(md.fsc.complete) <- rownames(md.fsc)
490 md.fsc.complete$row.names <- NULL
491 md.fsc.complete$Row.names <- NULL
492
493 # merged Phyloseq objects expects data to be in variable md.fsc.complete:
494 md.fsc.complete -> md.phyloseq

```

```

493 # garbage collection – save plots before this
494 rm(list = ls()[grep("^fil*", ls())])
495 rm(list = ls()[grep("^pca*", ls())])
496 rm(c.loc.group, c.sample.names, c.ss.loc.group, var.ind.touched, var.names)
497 rm(x.loc.group, x.sample.names, x.ss.loc.group)
498 rm(xc.loc.group, xc.ss.loc.group)
499 rm(md.fsc.complete)
500 rm(md,md.o.ss,md.fsc)
501
502 ##### 3 – Filtering and cleaning of phylotype data sets
503
504 # function – define new operator "not in"
505 '%!in%' <- function(x,y){ '%in%'(x,y)}
506
507 # function – Subtract Phylotypes contained in Australian controls, prune samples to PCM only
508 SubsetPCM <- function (ps_ob){
509
510     # OTU table – remove zero count OTUs
511     ps_ob.p <- prune_taxa(taxa_sums(ps_ob) > 0, ps_ob)
512
513     # isolate control samples – tested
514     ps_ob.aac <- prune_samples(sample_names(ps_ob.p)[c(grep("AACntrl", sample_names(ps_ob.p))], ps_ob.p )
515     ps_ob.inc <- prune_samples(sample_names(ps_ob.p)[c(grep("Inscntrl", sample_names(ps_ob.p))], ps_ob.p )
516     ps_ob.soc <- prune_samples(sample_names(ps_ob.p)[c(grep("soilcntrl", sample_names(ps_ob.p))], ps_ob.p )
517
518     # remove 0 count OTUs from subsets – tested
519     ps_ob.aac <- prune_taxa(taxa_sums(ps_ob.aac) > 0, ps_ob.aac)
520     ps_ob.inc <- prune_taxa(taxa_sums(ps_ob.inc) > 0, ps_ob.inc)
521     ps_ob.soc <- prune_taxa(taxa_sums(ps_ob.soc) > 0, ps_ob.soc)
522
523     # filter out Australian control phylotypes – tested
524     otu_table(ps_ob.p) <- subset(otu_table(ps_ob.p), subset = rownames(otu_table(ps_ob.p)) %!in% rownames(otu_table(ps_ob.inc)), drc
525     otu_table(ps_ob.p) <- subset(otu_table(ps_ob.p), subset = rownames(otu_table(ps_ob.p)) %!in% rownames(otu_table(ps_ob.soc)), drc
526
527     # keep only PCM samples
528     ps_ob.p <- prune_samples(sample_names(ps_ob.p)[c(grep("PCM", sample_names(ps_ob.p))], ps_ob.p )
529
530     # keep only PCM phylotypes
531     ps_ob.p <- prune_taxa(taxa_sums(ps_ob.p) > 0, ps_ob.p)
532
533

```

```

534 # name ranks properly for later clean up
535 colnames(tax_table(ps_ob.p))[1:7] <- c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")
536
537 # remove crap from taxonomy strings
538 tax_table(ps_ob.p) <- gsub("_", "", tax_table(ps_ob.p))
539 tax_table(ps_ob.p) <- gsub("-", "", tax_table(ps_ob.p))
540 tax_table(ps_ob.p) <- gsub(".", "", tax_table(ps_ob.p))
541 tax_table(ps_ob.p) <- gsub(" ", "", tax_table(ps_ob.p))
542
543 # remove zero count taxa again
544 ps_ob.p <- prune_taxa(taxa_sums(ps_ob.p) > 0, ps_ob.p)
545
546 # print summary
547 message(length(sample_names(ps_ob.p)), " Samples (", appendLF = FALSE)
548 message(100 * round(length(sample_names(ps_ob.p))/length(sample_names(ps_ob.p)), digits = 2), "%) and ", appendLF = FALSE)
549 message(length(taxa_names(ps_ob.p)), " Taxa (", appendLF = FALSE)
550 message(100 * round(length(taxa_names(ps_ob.p))/length(taxa_names(ps_ob.p)), digits = 2), "%) retained")
551
552 # return subsetted object
553 return(ps_ob.p)
554 }
555
556 # function - 18S clean up - * works only with 18S C97 T90 * -
557 GetClean18S <- function(clean.me){
558   assign("clean.me", clean.me, envir=globalenv())
559   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Arthropoda"),1] <-c("Ecdysozoa")
560   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Arthropoda"),2] <-c("Arthropoda")
561   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Nematoda"),1] <-c("Ecdysozoa")
562   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Nematoda"),2] <-c("Nematoda")
563   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Tardigrada"),1] <-c("Ecdysozoa")
564   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Tardigrada"),2] <-c("Tardigrada")
565   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Rotifera"),1] <-c("Lophotrochozoa")
566   tax_table(clean.me)[which(tax_table(clean.me)[,4] == "Rotifera"),2] <-c("Rotifera")
567   clean.me = subset_taxa(clean.me, Genus!="Insecta")
568   clean.me = subset_taxa(clean.me, Genus!="Maxillopoda")
569   clean.me = subset_taxa(clean.me, Genus!="Diplopoda")
570   tax_table(clean.me)[,3] <- NA
571   tax_table(clean.me)[,4] <- NA
572   tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Chromadorea"),3] <-c("Chromadorea")
573   tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Enoplea"),3] <-c("Enoplea")
574   tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Enoplea"),3] <-c("Enoplea")

```

```

575 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Chelicerata"),2] <-c("Chelicerata")
576 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Collembola"),3] <-c("Collembola")
577 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Arachnida"),3] <-c("Arachnida")
578 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Echiniscidae"),3] <-c("Heterotardigrada")
579 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Macrobiotidae"),3] <-c("Eutardigrada")
580 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Milnesiidae"),3] <-c("Eutardigrada")
581 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Sinantherina"),3] <-c("Monogononta")
582 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Sinantherina"),4] <-c("Flosculariacea")
583 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Sinantherina"),5] <-c("Flosculariidae")
584 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Cephalodella"),3] <-c("Monogononta")
585 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Cephalodella"),4] <-c("Ploimida")
586 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Cephalodella"),5] <-c("Notommatidae")
587 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Ploesoma"),3] <-c("Monogononta")
588 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Ploesoma"),4] <-c("Ploimida")
589 tax_table(clean.me)[which(tax_table(clean.me)[,6] == "Ploesoma"),5] <-c("Synchaetidae")
590 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Philodinidae"),3] <-c("Bdelloidea")
591 tax_table(clean.me)[which(tax_table(clean.me)[,5] == "Philodinidae"),4] <-c("Philodinida")
592 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,6]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,6]),6]
593 tax_table(clean.me)[grep("Plectidae",tax_table(clean.me)[,5]),4] <-c("Araeolaimida")
594 tax_table(clean.me)[grep("Rhabditidae",tax_table(clean.me)[,5]),4] <-c("Rhabditida")
595 tax_table(clean.me)[grep("Monhysteridae",tax_table(clean.me)[,5]),4] <-c("Monhysterida")
596 tax_table(clean.me)[grep("Cephalobidae",tax_table(clean.me)[,5]),4] <-c("Rhabditida")
597 tax_table(clean.me)[grep("Qudsiannematidae",tax_table(clean.me)[,5]),4] <-c("Dorylaimida")
598 tax_table(clean.me)[grep("Dolichodoridae",tax_table(clean.me)[,5]),4] <-c("Tylenchida")
599 tax_table(clean.me)[grep("Longidoridae",tax_table(clean.me)[,5]),4] <-c("Dorylaimida")
600 tax_table(clean.me)[grep("Echiniscidae",tax_table(clean.me)[,5]),4] <-c("Echiniscoidea")
601 tax_table(clean.me)[grep("Macrobiotidae",tax_table(clean.me)[,5]),4] <-c("Parachela")
602 tax_table(clean.me)[grep("Milnesiidae",tax_table(clean.me)[,5]),4] <-c("Apochela")
603 tax_table(clean.me)[grep("Nygolaimidae",tax_table(clean.me)[,5]),4] <-c("Dorylaimida")
604 tax_table(clean.me)[grep("Panagrolaimidae",tax_table(clean.me)[,5]),4] <-c("Rhabditida")
605 tax_table(clean.me)[grep("Xiphinematidae",tax_table(clean.me)[,5]),4] <-c("Dorylaimida")
606 tax_table(clean.me)[,6] <- tax_table(clean.me)[,7]
607 tax_table(clean.me)[,7] <- tax_table(clean.me)[,8]
608 tax_table(clean.me) <- tax_table(clean.me)[,-8]
609 tax_table(clean.me)[grep("Milnesium tardigradum",tax_table(clean.me)[,6]),6:7] <-c("Milnesium", "Milnesium tardigradum")
610 tax_table(clean.me)[grep("Plectidae sp. PDL 2005",tax_table(clean.me)[,6]),6:7] <-c(NA, NA)
611 tax_table(clean.me)[grep("uncultured nematode",tax_table(clean.me)[,6]),6:7] <-c(NA, NA)
612 tax_table(clean.me)[grep("Pseudechiniscus facettalis",tax_table(clean.me)[,6]),6:7] <-c("Pseudechiniscus", "Pseudechiniscus face")
613 tax_table(clean.me)[grep("Echiniscus testudo",tax_table(clean.me)[,6]),6:7] <-c("Echiniscus", "Echiniscus testudo")
614 tax_table(clean.me)[grep("Macrobiotus hufelandi group sp. NG 2008",tax_table(clean.me)[,6]),6:7] <-c("Macrobiotus", "Macrobiotus")
615 tax_table(clean.me)[grep("Paramacrobiotus richtersi group sp. NG 2008",tax_table(clean.me)[,6]),6:7] <-c("Paramacrobiotus", "Par

```

```

616 tax_table(clean.me)[grep("Halicephalobus", tax_table(clean.me)[,6]),6:7] <-c("Halicephalobus", "Halicephalobus cf. gingivalis")
617 tax_table(clean.me)[grep("Richtersius coronifer", tax_table(clean.me)[,6]),7] <-c("Richtersius coronifer")
618 tax_table(clean.me)[grep("Richtersius coronifer", tax_table(clean.me)[,7]),6] <-c("Richtersius")
619 tax_table(clean.me)[grep("Sminthuridae", tax_table(clean.me)[,6]),4] <-c("Symphypleona")
620 tax_table(clean.me)[grep("Symphypleona", tax_table(clean.me)[,4]),5] <-c("Sminthuridae")
621 tax_table(clean.me)[grep("Symphypleona", tax_table(clean.me)[,4]),6:7] <-c(NA)
622 tax_table(clean.me)[grep("Hexapoda", tax_table(clean.me)[,5]),5:7] <-c(NA)
623 tax_table(clean.me)[grep("Damon", tax_table(clean.me)[,6]),1:5] <-c("Ecdysozoa", "Chelicerata", "Arachnida", "Amblypygi", "Phryni")
624 tax_table(clean.me)[grep("Chelicerata", tax_table(clean.me)[,5]),5] <-c(NA)
625 tax_table(clean.me)[grep("Meristolohmannia", tax_table(clean.me)[,6]),1] <-c("Ecdysozoa")
626 tax_table(clean.me)[grep("Meristolohmannia", tax_table(clean.me)[,6]),2] <-c("Chelicerata")
627 tax_table(clean.me)[grep("Meristolohmannia", tax_table(clean.me)[,6]),3] <-c("Arachnida")
628 tax_table(clean.me)[grep("Meristolohmannia", tax_table(clean.me)[,6]),4] <-c("Oribatida")
629 tax_table(clean.me)[grep("Meristolohmannia", tax_table(clean.me)[,6]),5] <-c("Lohmanniidae")
630 tax_table(clean.me)[grep("Hoplotaimidae", tax_table(clean.me)[,5]),4] <-c("Tylenchida")
631 unique(tax_table(clean.me))[order(unique(tax_table(clean.me))[,4])]
632 result <- clean.me
633 rm("clean.me", envir=globalenv())
634 return(result)
635 }
636
637 # function - COI clean up - * works only with COI C97 T75 * -
638 GetCleanCOI <- function(clean.me){
639   assign("clean.me", clean.me, envir=globalenv())
640   # clean to class level:
641   clean.me = subset_taxa(clean.me, Family!="Myriapoda")
642   clean.me = subset_taxa(clean.me, Family!="Hexapoda")
643   tax_table(clean.me)[,1] <- tax_table(clean.me)[,3]
644   tax_table(clean.me)[,2] <- tax_table(clean.me)[,4]
645   tax_table(clean.me)[,3] <- tax_table(clean.me)[,5]
646   tax_table(clean.me)[,4] <- tax_table(clean.me)[,6]
647   tax_table(clean.me) <- tax_table(clean.me)[,-15:-19]
648   tax_table(clean.me)[grep("Chelicerata", tax_table(clean.me)[,3]),2] <-c("Chelicerata")
649   tax_table(clean.me)[grep("Arachnida", tax_table(clean.me)[,4]),3] <-c("Arachnida")
650
651   # clean order level:
652   tax_table(clean.me)[,4:5] <-(NA)
653   tax_table(clean.me)[grep("Arachnida", tax_table(clean.me)[,6]),6] <-c(NA)
654   tax_table(clean.me)[grep("Adinetida", tax_table(clean.me)[,6]),4] <-c("Adinetida")
655   tax_table(clean.me)[grep("Araeolaimida", tax_table(clean.me)[,6]),4] <-c("Araeolaimida")
656   tax_table(clean.me)[grep("Tylenchida", tax_table(clean.me)[,6]),4] <-c("Tylenchida")

```

```
657 tax_table(clean.me)[grep("Dorylaimia",tax_table(clean.me)[,6]),4] <-c("Dorylaimida")
658 tax_table(clean.me)[grep("Araneae",tax_table(clean.me)[,7]),4] <-c("Araneae")
659 tax_table(clean.me)[grep("Sarcoptiformes",tax_table(clean.me)[,9]),4] <-c("Sarcoptiformes")
660 tax_table(clean.me)[grep("Prostigmata",tax_table(clean.me)[,10]),4] <-c("Trombidiformes")
661
662 # clean Family and Genus Level
663 tax_table(clean.me)[,5:7] <- (NA)
664 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,8]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,8]),8]
665 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,9]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,9]),9]
666 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,10]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,10]),10]
667 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,11]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,11]),11]
668 tax_table(clean.me)[grep("idae",tax_table(clean.me)[,13]),5] <- tax_table(clean.me)[grep("idae",tax_table(clean.me)[,13]),13]
669 tax_table(clean.me)[grep("Adineta",tax_table(clean.me)[,8]),6] <-c("Adineta")
670 tax_table(clean.me)[grep("Adineta",tax_table(clean.me)[,6]),7] <-c("Adineta vaga")
671 tax_table(clean.me)[grep("Adineta",tax_table(clean.me)[,6]),5] <-c("Adinetidae")
672 tax_table(clean.me)[grep("Plectus",tax_table(clean.me)[,9]),6] <-c("Plectus")
673 tax_table(clean.me)[grep("Plectus murrayi",tax_table(clean.me)[,10]),7] <-c("Plectus murrayi")
674 tax_table(clean.me)[grep("Plectus cf frigophilus",tax_table(clean.me)[,10]),7] <-c("Plectus cf frigophilus")
675
676 # clean species level
677 tax_table(clean.me)[grep("Bursaphelenchus cocophilus",tax_table(clean.me)[,11]),7] <-c("Bursaphelenchus cocophilus")
678 tax_table(clean.me)[grep("Bursaphelenchus cocophilus",tax_table(clean.me)[,11]),6] <-c("Bursaphelenchus")
679 tax_table(clean.me)[grep("Antistea brunnea",tax_table(clean.me)[,13]),7] <-c("Antistea brunnea")
680 tax_table(clean.me)[grep("Antistea brunnea",tax_table(clean.me)[,13]),6] <-c("Antistea")
681 tax_table(clean.me)[grep("Arctosa insignita",tax_table(clean.me)[,13]),7] <-c("Arctosa insignita")
682 tax_table(clean.me)[grep("Arctosa insignita",tax_table(clean.me)[,13]),6] <-c("Arctosa")
683 tax_table(clean.me)[grep("Pardosa lapponica",tax_table(clean.me)[,13]),7] <-c("Pardosa lapponica")
684 tax_table(clean.me)[grep("Pardosa lapponica",tax_table(clean.me)[,13]),6] <-c("Pardosa")
685 tax_table(clean.me)[grep("Robertus lyriifer",tax_table(clean.me)[,13]),7] <-c("Robertus lyriifer")
686 tax_table(clean.me)[grep("Robertus lyriifer",tax_table(clean.me)[,13]),6] <-c("Robertus")
687 tax_table(clean.me)[grep("Erigone tirolensis",tax_table(clean.me)[,14]),7] <-c("Erigone tirolensis")
688 tax_table(clean.me)[grep("Erigone tirolensis",tax_table(clean.me)[,14]),6] <-c("Erigone")
689 tax_table(clean.me)[grep("Estrandia grandaeva",tax_table(clean.me)[,14]),7] <-c("Estrandia grandaeva")
690 tax_table(clean.me)[grep("Estrandia grandaeva",tax_table(clean.me)[,14]),6] <-c("Estrandia")
691 tax_table(clean.me)[grep("Meioneta simplex",tax_table(clean.me)[,14]),7] <-c("Meioneta simplex")
692 tax_table(clean.me)[grep("Meioneta simplex",tax_table(clean.me)[,14]),6] <-c("Meioneta")
693 tax_table(clean.me)[grep("Xiphinema",tax_table(clean.me)[,11]),6] <-c("Xiphinema")
694
695 # erase unneeded columns
696 tax_table(clean.me) <- tax_table(clean.me)[,-8:-14]
697
```



```

698 # return object
699 result <- clean.me
700 rm("clean.me", envir=globalenv())
701 return(result)
702 }
703
704 # function - remove 0 count Phylotypes and Samples
705 RmZeroCounts <- function(ps_obj){
706
707   # filter Phylotypes
708   ps_obj <- prune_taxa(taxa_sums(ps_obj) > 0, ps_obj)
709
710   # filter samples
711   ps_obj <- prune_samples(sample_sums(ps_obj) > 0, ps_obj)
712
713   # return object
714   return(ps_obj)
715 }
716
717 # function - plot phyla across samples, requires two Phyloseq objects
718 PlotUnmerged <- function(ps_obj1, ps_obj2, level = c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")){
719   plot.e <- plot_bar(ps_obj1, fill=level, facet_grid="Phylum~Location") + theme_bw() + theme(axis.text.x = element_text(angle = 90,
720   plot.c <- plot_bar(ps_obj2, fill=level, facet_grid="Phylum~Location") + theme_bw() + theme(axis.text.x = element_text(angle = 90,
721   grid.arrange(plot.e, plot.c, nrow=2)
722   }
723
724 ## a - subset datasets to PCM data only, substract and remove Australian controls, remove Antarctic controls
725
726 ps.e <- SubsetPCM(ps.e)
727 ps.c <- SubsetPCM(ps.c)
728
729 ## b - Gene-specific cleanup of taxonomy tables - remove insects - remove Arthropods - remove RH regions
730
731 ps.e <- GetClean18S(ps.e)
732 ps.c <- GetCleanCOI(ps.c)
733
734 # remove RH region
735 ps.e <- subset_samples(ps.e, sample_data(ps.e)$Location != "Reinbolt_Hills" )
736 ps.c <- subset_samples(ps.c, sample_data(ps.c)$Location != "Reinbolt_Hills" )
737
738 # remove Arthropods

```

```

739   ps.e <- subset_taxa(ps.e, tax_table(ps.e)[,2] != "Arthropoda" )
740   ps.c <- subset_taxa(ps.c, tax_table(ps.c)[,2] != "Arthropoda" )
741
742   ## c - Remove empty samples to increase data density
743
744   ps.c <- RmZeroCounts(ps.c)
745   ps.e <- RmZeroCounts(ps.e)
746
747   ## d - Check finished unmerged objects
748
749   PlotUnmerged(ps.e, ps.c, "Phylum")
750
751   ##### 4 - merge Phylotype data and sample data into one object
752
753   # function: rename Phylotype components for merging, do tip agglomeration of Phylogenetic trees to reduce distance between samples
754   PrepMerge <- function(ps.e, ps.c, tipglom) {
755
756     # Check input
757     if(is.null(tipglom)) {
758       x <- 0
759       warning('Cluster parameter not set. Tip agglomeration will be omitted.')
760     }
761
762     ## rename phylotype identifiers in both objects
763     # OTU tables
764     dimnames(ps_e@otu_table@Data)[[1]] <- paste0("ets_", dimnames(ps_e@otu_table@Data)[[1]])
765     dimnames(ps_c@otu_table@Data)[[1]] <- paste0("coi_", dimnames(ps_c@otu_table@Data)[[1]])
766
767     # taxonomy table
768     dimnames(ps_e@tax_table@Data)[[1]] <- paste0("ets_", dimnames(ps_e@tax_table@Data)[[1]])
769     dimnames(ps_c@tax_table@Data)[[1]] <- paste0("coi_", dimnames(ps_c@tax_table@Data)[[1]])
770
771     # reference sequences table
772     ps_e@refseq@ranges@NAMES <- paste0("ets_", ps_e@refseq@ranges@NAMES)
773     ps_c@refseq@ranges@NAMES <- paste0("coi_", ps_c@refseq@ranges@NAMES)
774
775     # tree tip labels
776     ps_e@phy_tree$tip.label <- paste0("ets_", ps_e@phy_tree$tip.label)
777     ps_c@phy_tree$tip.label <- paste0("coi_", ps_c@phy_tree$tip.label)
778
779     message("Renamed Phylotypes")

```

```

# OTU table: Phylotype identifiers
# OTU table: Phylotype identifiers

```

```

# TAX table: Phylotype identifiers
# TAX table: Phylotype identifiers

```

```

# SEQS - reference sequences - Phylotype identifiers
# SEQS - reference sequences - Phylotype identifiers

```

```

# TREE labels: Phylotype identifiers - add prefix
# TREE labels: Phylotype identifiers - add prefix

```

```

780 ## rename sample identifiers in both objects
781 # OTU tables
782 dimnames(ps_e@otu_table@Data)[[2]] <- gsub('.{3}$', '', substring(dimnames(ps_e@otu_table@Data)[[2]],5)) # OTU table: Sample names
783 dimnames(ps_c@otu_table@Data)[[2]] <- gsub('.{3}$', '', substring(dimnames(ps_c@otu_table@Data)[[2]],5)) # OTU table: Sample names
784
785 # Sample data
786 ps_e@sam_data@row.names <- gsub('.{3}$', '', substring(ps_e@sam_data@row.names,5)) # SAMPLE data: Sample names - strip prefixes
787 ps_c@sam_data@row.names <- gsub('.{3}$', '', substring(ps_c@sam_data@row.names,5)) # SAMPLE data: Sample names - strip prefixes
788
789 message("Renamed Samples")
790
791 ## Tree plotting and tip agglomeration
792 # show trees and number of taxa - original
793 message("Generating original trees")
794 p1 <- plot_tree(ps_e, color = "Location", shape = "Phylum", ladderize = "left", size = "abundance", base.spacing = 0.03) +
795   ggtitle("18S Phylotypes (not agglomerated)") + theme_bw() +
796   theme(axis.title.x = element_blank(), axis.text.x = element_blank()) +
797   ylab(paste("Original number of taxa:", ntaxa(ps_e)))
798
799 p2 <- plot_tree(ps_c, color = "Location", shape = "Phylum", ladderize = "left", size = "abundance", base.spacing = 0.03) +
800   ggtitle("COI Phylotypes (not agglomerated)") +
801   theme_bw() + theme(axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
802   ylab(paste("Original number of taxa:", ntaxa(ps_c)))
803
804 message("Generating trimmed trees")
805 # tip agglomeration to reduce noise
806 if(tipglom > 0 & tipglom < 1){
807   ps_c <- tip_glom(ps_c, h = tipglom)
808   ps_e <- tip_glom(ps_e, h = tipglom)
809 }
810
811 # show trees and number of taxa - original
812 p3 <- plot_tree(ps_e, color = "Location", shape = "Phylum", ladderize = "left", size = "abundance", base.spacing = 0.03) +
813   ggtitle(paste0("18S Phylotypes (cut height ", tipglom, ")")) + theme_bw() +
814   theme(axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
815   ylab(paste("Retained number of taxa for 18S:", ntaxa(ps_e)))
816
817 p4 <- plot_tree(ps_c, color = "Location", shape = "Phylum", ladderize = "left", size = "abundance", base.spacing = 0.03) +
818   ggtitle(paste0("COI Phylotypes (cut height ", tipglom, ")")) + theme_bw() +
819   theme(axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
820   ylab(paste("Retained number of taxa for COI:", ntaxa(ps_c)))

```

```
821 # plot out trees
822 grid.arrange(p1, p2, p3, p4, ncol=2)
823
824 # store results in list
825 PrepResults <- list(ps_e, ps_c)
826
827 # return list result
828 return(PrepResults)
829
830 }
831
832 ## a - prepare merging
833
834 # save copy of object prior to tip glom
835 ps.c.preTipGlom <- ps.c
836 ps.e.preTipGlom <- ps.e
837 rm(ps.c)
838 rm(ps.e)
839
840 # apply tip glom function
841 ModPSinList <- PrepMerge(ps.e.preTipGlom, ps.c.preTipGlom, 0.1)
842
843 # store output - and collect garbage
844 ps.e.postTipGlom <- ModPSinList[[1]]
845 ps.c.postTipGlom <- ModPSinList[[2]]
846 rm(ModPSinList)
847
848 # filter, only for testing
849 ps.e.postTipGlom <- RmZeroCounts(ps.e.postTipGlom)
850 ps.c.postTipGlom <- RmZeroCounts(ps.c.postTipGlom)
851
852 ## b - merge OTU and TAX tables
853
854 # merge OTU tables - using list elements from previous step
855 merged.otu.tab <- merge_phyloseq_pair(otu_table(ps.e.postTipGlom), otu_table(ps.c.postTipGlom))
856
857 # merge TAX tables - using list elements from previous step
858 merged.tax.tab <- merge_phyloseq_pair(tax_table(ps.e.postTipGlom), tax_table(ps.c.postTipGlom))
859
860 # merge REF sequences - using list elements from previous step
861 merged.ref.seqs <- merge_phyloseq_pair(refseq(ps.e.postTipGlom), refseq(ps.c.postTipGlom))
```

Saved: 18/06/15 16:11:14

```

862 ## c - merge *subsetting / preprocessed* observation metadata into sample data
863
864 # this is scaled and centered soil and xrd data
865 merged.sample.data <- sample_data(md.phyloseq)
866
867 ## d - create Phyloseq object - and collect garbage
868
869 # merge objects
870 ps.merged <- merge_phyloseq(merged.otu.tab, merged.tax.tab, merged.ref.seqs, merged.sample.data)
871
872 # clean
873 ps.merged <- RmZeroCounts(ps.merged)
874
875 # rename lat long columns in Phyloseq object to be recognized by Phylogeo
876 names(sample_data(ps.merged))[names(sample_data(ps.merged))=="g_latitude"] <- "Latitude"
877 names(sample_data(ps.merged))[names(sample_data(ps.merged))=="g_longitude"] <- "Longitude"
878
879 # copy final data
880 MergedPS <- ps.merged
881
882 # collect garbage
883 rm(merged.otu.tab, merged.tax.tab, merged.ref.seqs, merged.sample.data, ps.merged)
884
885 ##### 5 - functions for Analysis
886
887 ## a - Simple maps for testing
888 SimplyMapSamples <- function(ps_ob){
889   map_phyloseq(ps_ob, region="Antarctica", color="Location", jitter=TRUE, jitter.x=2,jitter.y=2, ) +
890     coord_map("ortho", orientation=c(-90, 0, 0)) + geom_path()
891 }
892
893 ## b - Simple NMDS for testing
894 SimplyNMDSplot <- function(ps_ob, tax_level = c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")){
895   # ordinate
896   ord <- ordinate(ps_ob, "NMDS", "jaccard")
897   # plot ordination
898   plot_ordination(ps_ob, ord, type = "split", shape = "Location", color = tax_level, title = "NMDS(Bray) - Location vs. Classes") + then

```

```

903 }
904
905 ## c - Taxon agglomeration - know what you put in here first!
906 TaxGlom <- function(ps_ob, tax_table_column){
907
908   # Warning message
909   message(paste("Look at taxonomy table first so that you know what you will loose"))
910
911   # Taxon agglomeration at defined level, defaults to Superphylum
912   return(phyloseq::tax_glom(ps_ob, taxrank=phyloseq::rank_names(ps_ob)[tax_table_column]))
913 }
914
915 ## d - Plot Samples based on location
916 # CAUTION: Uses global env
917 PlotMergedSamples <- function(ps_ob) {
918
919   # Initialize function - REQUIRES GARBAGE COLLECTION!
920   assign("ps_ob", ps_ob, envir=globalenv())
921   message(paste("Check Garbage Collection after this step. Object ps_ob is being written to global Environment"))
922
923
924   # subset by location
925   ps_ob.lt <- subset_samples(ps_ob, sample_data(ps_ob)[,which(colnames(sample_data(ps_ob)) == "Location")]) == "Lake_Terrasovoje")
926   ps_ob.me <- subset_samples(ps_ob, sample_data(ps_ob)[,which(colnames(sample_data(ps_ob)) == "Location")]) == "Mawson_Escarpment")
927   ps_ob.mm <- subset_samples(ps_ob, sample_data(ps_ob)[,which(colnames(sample_data(ps_ob)) == "Location")]) == "Mount_Menzies")
928
929   # remove empty samples and rows
930   ps_ob.lt <- RmZeroCounts(ps_ob.lt)
931   ps_ob.me <- RmZeroCounts(ps_ob.me)
932   ps_ob.mm <- RmZeroCounts(ps_ob.mm)
933
934   # normalize sample counts
935   ps_ob.lt <- transform_sample_counts(ps_ob.lt, function(OTU) OTU/sum(OTU))
936   ps_ob.me <- transform_sample_counts(ps_ob.me, function(OTU) OTU/sum(OTU))
937   ps_ob.mm <- transform_sample_counts(ps_ob.mm, function(OTU) OTU/sum(OTU))
938
939   # create plots
940   p1 <- plot_bar(ps_ob.mm, fill="Class", facet_grid="Phylum~Location") + theme_bw() + theme(axis.text.x = element_text(angle = 90, hjust
941   p2 <- plot_bar(ps_ob.me, fill="Class", facet_grid="Phylum~Location") + theme_bw() + theme(axis.text.x = element_text(angle = 90, hjust
942   p3 <- plot_bar(ps_ob.lt, fill="Class", facet_grid="Phylum~Location") + theme_bw() + theme(axis.text.x = element_text(angle = 90, hjust
943

```

```

944 # combine plots
945 grid.arrange(p1, p2, p3, nrow = 3)
946
947 # garbage cleaning
948 rm(ps_ob, ps_ob.lt, ps_ob.me, ps_ob.mm, p1, p2, p3 )
949 rm(ps_ob, envir=globalenv())
950
951 }
952
953 ## e – Subset to Taxonomic Level
954 # CAUTION: Uses global env
955 IsolateTaxa <- function(ps, level = c("Superphylum", "Phylum", "Class", "Order", "Family", "Genus", "Species")){
956
957   # Initialize function – REQUIRES GARBAGE COLLECTION!
958   require("phyloseq")
959   message(paste("Check Garbage Collection after this step. Multiple objects / variables are being written to global Environment"))
960   assign("ps_ob", ps, envir=globalenv())
961
962   # tax table index number corresponding to specified string
963   level.index <- which(colnames(tax_table(ps_ob)) == level)
964
965   # vector with unique taxa at specified level
966   unq.taxa.lev <- c(unique(tax_table(ps_ob)[,which(colnames(tax_table(ps_ob)) == level)]))
967
968   # initialize list (length + 1) to carry through one not – subsetted object
969   ps.list <- vector("list", length(unq.taxa.lev) + 1)
970
971   # name list components with "All taxa" + taxa names in taxa vector
972   names(ps.list) <- c(unq.taxa.lev, "All_Taxa" )
973
974   # store complete object at list position 1
975   ps.list[[length(unq.taxa.lev) + 1]] <- RmZeroCounts(ps_ob)
976
977   #stuff to do the number of times that unq.taxa.lev is long
978   for(i in 1:length(unq.taxa.lev)){
979     # print the taxon string that is needed for pruning OTU table
980     message(paste("Subsetting to", level, unq.taxa.lev[i]))
981
982     # working around subset_taxa() 1/2
983     assign("level.index", level.index, envir=globalenv())
984     assign("unq.taxa.lev", unq.taxa.lev[i], envir=globalenv())

```

```
985 # subset the OTU table of the input phyloseq object and store in list after the first item
986 ps.list[[i]] <- RmZeroCounts(subset_taxa(ps_ob, tax_table(ps_ob)[,level.index] == unq.taxa.lev))
987
988
989 # working around subset_taxa() 2/2
990 rm("level.index", envir=globalenv())
991 rm("unq.taxa.lev", envir=globalenv())
992
993 }
994
995 # Finishing up
996 results <- ps.list
997
998 # Garbage collection after testing
999 rm(level, level.index, unq.taxa.lev, ps.list, i)
1000 rm(ps_ob, envir=globalenv())
1001
1002 return(results)
1003
1004 }
1005
1006 ## f - Wrapper function to convert OTU tables for Vegan
1007 vOTU <- function(ps) {
1008   OTU <- otu_table(ps)
1009   if (taxa_are_rows(OTU)) {
1010     OTU <- t(OTU)
1011   }
1012   return(as(OTU, "matrix"))
1013 }
1014
1015 ## g - Calculate distance matrices - return list
1016 # Creates binary jaccard distance matrices on list elements, returns matrices in list
1017 VegDist <- function(ps){
1018   # calculate matrix
1019   dis <- vegdist(vOTU(ps), "jaccard", binary=TRUE)
1020   # dis <- vegdist(vOTU(ps), "jaccard")
1021
1022   # return matrix
1023   return(dis)
1024 }
1025
```



```

1026 }
1027
1028 ## h - calculate MDS of biotic data
1029 # Creates MDS objects of list elements - returns list
1030 Mds <- function(ps){
1031
1032     # do metaMDS analysis
1033     mds <- try(metaMDS(vOTU(ps), "bray", binary=TRUE, trymax = 100))
1034     # mds <- try(metaMDS(vOTU(ps), "jaccard", trymax = 250))
1035
1036     # return MDS
1037     return(mds)
1038 }
1039
1040 ## i - Get Observation and Species data
1041 # returns list
1042 GetSpecObs <- function(ps_ob){
1043
1044     # Species data
1045     spec.df <- data.frame(vOTU(ps_ob))
1046
1047     # Observation data
1048     sobs.df <- data.frame(sample_data(ps_ob))
1049
1050     # define variables that are to be used
1051     var.names <- c("Gravel", "Texture", "Ammonium", "Nitrate", "Phosphorus", "Potassium", "Sulphur",
1052                   "Org_Carbon", "pH_H2O", "Quartz", "Feltspar", "Titanite", "Pyr_Amp_Gar", "Micas",
1053                   "Dolomite", "Kao_Chlor", "Calcite", "Chlorite", "Elevation", "Slope", "Soil_Temp")
1054
1055     # select sample data column indices that are to be used for preprocessing
1056     var.ind.touched <- which(names(sobs.df) %in% var.names)
1057
1058     # extract clomuns
1059     sobs.df <- sobs.df[,var.ind.touched]
1060
1061     # jut in case:
1062     obs <- data.frame(as.matrix(sobs.df))
1063     spc <- data.frame(as.matrix(spec.df))
1064
1065
1066

```

Saved: 18/06/15 16:11:14

```
1067 # return final models
1068 return(list(spc, obs))
1069
1070 }
1071
1072 ## k - Return Models 1 and 0
1073 ReturnModels <- function(obs_list, analysis = c("cca", "rda", "cap")){
1074
1075   spc <- obs_list[[1]]
1076   obs <- obs_list[[2]]
1077
1078   # select and do analysis
1079   if(analysis == "cca"){
1080     message(paste("Doing CCA"))
1081
1082     mod1 <- try(vegan::cca(spc ~ ., obs))
1083     mod0 <- try(vegan::cca(spc ~ 1, obs))
1084
1085   }
1086   if(analysis == "rda"){
1087     message(paste("Doing RDA"))
1088
1089     mod1 <- try(vegan::rda(spc ~ ., obs))
1090     mod0 <- try(vegan::rda(spc ~ 1, obs))
1091
1092   }
1093   if(analysis == "cap"){
1094     message(paste("Doing Capscale"))
1095
1096     mod1 <- try(vegan::capscale(spc ~ ., obs))
1097     mod0 <- try(vegan::capscale(spc ~ 1, obs))
1098
1099   }
1100
1101   # return final models
1102   return(list(mod1, mod0))
1103
1104 }
1105
1106 }
1107
```

Saved: 18/06/15 16:11:14

```
1108 ## l - Get Models with best AIC score for Constrained ordinations
1109 # CAUTION: Uses global env
1110 ModStep <- function(model_list, data_list){
1111
1112   message(paste("Check Garbage Collection after this step. Multiple objects / variables are being written to global environment. Function
1113   require("phyloseq", quietly = TRUE, warn.conflicts = FALSE)
1114
1115   assign("mod1", model_list[[1]], envir=globalenv())
1116   assign("mod0", model_list[[2]], envir=globalenv())
1117
1118   assign("spc", data_list[[1]], envir=globalenv())
1119   assign("obs", data_list[[2]], envir=globalenv())
1120
1121   if(model_list[[1]]$method == "cca"){
1122
1123     # Message
1124     message(paste("Model searching for CCA"))
1125
1126     # necessary for step function and cca
1127     try(detach("package:phylogeo"))
1128     try(detach("package:phyloseq"))
1129     try(detach("package:ade4"))
1130
1131
1132     # step from unconstrained to maximum constrained model
1133     mod <- try(stats::step(mod0, scope = formula(mod1), test = "perm"))
1134
1135     # reload so that the other function work again
1136     require("phylogeo", quietly = TRUE, warn.conflicts = FALSE)
1137     require("phyloseq", quietly = TRUE, warn.conflicts = FALSE)
1138     require("ade4", quietly = TRUE, warn.conflicts = FALSE)
1139
1140   }
1141   if(model_list[[1]]$method == "rda"){
1142
1143     # Message
1144     message(paste("Model searching for RDA"))
1145
1146     # step from unconstrained to maximum constrained model
1147     mod <- try(stats::step(mod0, scope = formula(mod1), test = "perm"))
1148
```

```

1149 }
1150
1151 if(model_list[[1]]$method == "capscale"){
1152
1153
1154   # Message
1155   message(paste("Model searching for CAP"))
1156
1157   # step from unconstrained to maximum constrained model
1158   mod <- try(stats::step(mod0, scope = formula(mod1), test = "perm"))
1159
1160 }
1161
1162 rm("mod1", envir=globalenv())
1163 rm("mod0", envir=globalenv())
1164
1165 rm("spc", envir=globalenv())
1166 rm("obs", envir=globalenv())
1167
1168 return(mod)
1169
1170 }
1171
1172 ## m - Plot CA items in lists, needs Models in List and List with Phyloseq objects and some other things
1173 PlotCA <- function(mod_names, mod_list, ps_names, ps_list, taxa_level, site_label, text_labels, dim1, dim2){
1174   # warning message
1175   message(paste("Check axes!"))
1176   # check input data
1177   if(mod_names != ps_names){stop("List names do not match. Check input data.")}
1178   # suppress failed model calculation
1179   if(class(mod_list) != "try-error"){
1180     # create empty plot
1181     try(p <- plot(mod_list, choices = c(dim1, dim2), type="n", scaling = 3))
1182
1183     # modify sample labels
1184     try(rownames(p$sites) <- suppressWarnings(t(sample_data(ps_list)[which(rownames(p$sites) %in% rownames(sample_data(ps_list)
1185     # modify species labels
1186     try(rownames(p$species) <- tax_table(ps_list)[which(rownames(p$species) %in% rownames(tax_table(ps_list))), taxa_level])
1187
1188     # add environmental vectors - via original model
1189     try(points(mod_list, choices = c(dim1, dim2), display = "bp", scaling = 3, col = "blue"))

```

```

1190 try(text(mod_list, choices = c(dim1, dim2), display = "bp", scaling = 3, cex = 0.8, col = "blue"))
1191
1192 # add site and species labels via modified plot object
1193 try(points(p, "sites", pch=21, col="black", bg="grey"))
1194 try(points(p, "species", pch=21, col="red", bg="yellow"))
1195
1196
1197 # add text labels if desired:
1198 try(if(text_labels == TRUE){
1199     text(p, "sites", col="black", cex=1, offset = 1)
1200 })
1201 try(text(p, "species", col="red", cex=1, offset = 1))
1202
1203 # add title
1204 try(title(paste0(mod_names, " (", length(rownames(p$species)), " taxa", length(rownames(p$sites)), " samples )"))
1205
1206 # interactive
1207 # message(paste("Choose important sites by clicking. Hit ESC to continue"))
1208 # identify(p, "sites", labels = c("Mount_Menzies", "Mawson_Escarpment", "Lake_Terrasovoje"))
1209
1210 }
1211 }
1212
1213 ## n - Get List with subsetted to taxa pairs
1214 TrimTaxaPairs <- function(ps_list){
1215
1216     # Initialize function - REQUIRES GARBAGE COLLECTION!
1217     # message(paste("Check Garbage Collection after this step. Multiple objects / variables are being written to global Environment"))
1218     # assign("ps_ob", ps, envir=globalenv())
1219     # ps_list <- PhyseqList
1220
1221     # Erase liste item "All_Taxa"
1222     ps_list$All_Taxa <- NULL
1223
1224     # Store taxa to subset in vector
1225     tax_vector <- names(ps_list)
1226
1227     # Initialize List, list position, names vector
1228     results <- vector("list", length(tax_vector) * (length(tax_vector) -1))
1229     list_pos = 1
1230     list_names <- vector('character')

```

```
1231 # Step through tax vector
1232   for(i in 1:length(tax_vector)){
1233
1234
1235     # Step through list items
1236     for(j in 1:length(names(ps_list))){
1237
1238       # exclude matching pairs
1239       if(tax_vector[i] != names(ps_list)[j]){
1240
1241         # store objects for modification
1242         ps_to_trim <- (ps_list)[[i]]
1243         ps_trimmed <- (ps_list)[[i]]
1244         ps_for_trim <- (ps_list)[[j]]
1245
1246         # subset samples
1247         ps_trimmed <- prune_samples(sample_names(ps_to_trim) %in% sample_names(ps_for_trim), ps_trimmed)
1248
1249         # print diagnostics
1250         message(paste0(names(ps_list)[i], " to trim: ", nsamples(ps_to_trim), "; ", names(ps_list)[j], " for trim
1251
1252         # store result in list
1253         results[[list_pos]] <- ps_trimmed
1254         list_names <- append(list_names, paste0( names(ps_list)[i], "_by_", names(ps_list)[j]) )
1255
1256         # increase list counter
1257         list_pos = list_pos + 1
1258
1259       }
1260
1261     }
1262
1263   }
1264
1265   # name list items
1266   names(results) <- list_names
1267
1268   # garbage collection
1269   rm(ps_list, list_pos, list_names, tax_vector, ps_to_trim, ps_trimmed, ps_for_trim)
1270
1271   return(results)
```

```
1272 }
1273
1274 ## o - Plot MDS - could be combined with Plot CA with little work
1275 PlotMDS <- function(mds_ob, ps_ob, site_label = "Location", taxa_level = 2, text_labels = TRUE){
1276
1277   mds_plot <- plot(mds_ob, type = "n")
1278
1279   # modify sample labels
1280   rownames(mds_plot$sites) <- suppressWarnings(t(sample_data(ps_ob)[which(rownames(mds_plot$sites) %in% rownames(sample_data(ps_ob
1281
1282   # modify species labels
1283   rownames(mds_plot$species) <- tax_table(ps_ob)[which(rownames(mds_plot$species) %in% rownames(tax_table(ps_ob))), taxa_level]
1284
1285   # add site and species labels via modified plot object
1286   points(mds_plot, "sites", pch=21, col="black", bg="grey")
1287   points(mds_plot, "species", pch=21, col="red", bg="yellow")
1288
1289
1290   # add text labels if desired:
1291   if(text_labels == TRUE){
1292     text(mds_plot, "sites", col="black", cex=1, offset = 5)
1293     text(mds_plot, "species", col="red", cex=1, offset = 5)
1294
1295
1296     # add title
1297     title(paste0("NMDS: ", length(rownames(mds_plot$species)), " taxa", length(rownames(mds_plot$sites)), " samples"))
1298
1299   }
1300
1301 ##### 6 - Constrained ordinations: CCA / RDA / Capscale
1302
1303 ## a - Tax glom at class level
1304 MergedPSGlom <- TaxGlom(MergedPS, 3)
1305
1306 ## b - plot Phyloseq object and also do a PCA, to see what looks better
1307
1308 # plot out before taxon agglomeration
1309 PlotMergedSamples(MergedPS)
1310
1311 # tax table - used in manuscript as SI table six
```

```
1313 tax_table(MergedPS)
1314 length(unique(tax_table(MergedPS)[,1]))
1315 length(unique(tax_table(MergedPS)[,2]))
1316 length(unique(tax_table(MergedPS)[,3]))
1317 length(unique(tax_table(MergedPS)[,4]))
1318 length(unique(tax_table(MergedPS)[,5]))
1319 length(unique(tax_table(MergedPS)[,6]))
1320 length(unique(tax_table(MergedPS)[,7]))
1321
1322 # plot out after taxon agglomeration – used in manuscript
1323 PlotMergedSamples(MergedPSGlom)
1324
1325 # PCA on taxon-agglomerated object
1326
1327 # Tax glom at class level, or order level (4)
1328 MergedPSGlomPCA <- TaxGlom(MergedPS, 3)
1329
1330 # get OTU table as transposed matrix
1331 otus <- data.frame(vOTU(MergedPSGlomPCA))
1332
1333 # name OTUs correctly
1334 colnames(otus) <- tax_table(MergedPSGlomPCA)[which(colnames(otus) %in% rownames(tax_table(MergedPSGlomPCA)))], 3]
1335
1336 # summarize raw data
1337 summary(otus) # 103, 1
1338
1339 # pre-process abundance values
1340 otus.pp <- Trans(otus)
1341
1342 # summarize transformation
1343 summary(otus.pp)
1344
1345 # calculate PCs on correlation matrices
1346 pca.otus.pp <- prcomp(otus.pp, center = TRUE, scale = TRUE) # PCA on transformed phyloseq OTU table
1347 pca.otus.pp
1348
1349 # get variance plots
1350 plot(pca.otus.pp, type = "l")
1351
1352 # inspect original PCA objects (transformed, scaled cntred, but with all original PCs)
1353 pca.otus.pp$rotation
```



```

1354
1355 # check information content using caret package – retains PCs with 95% variation content – see objects
1356 trans_otu <- preProcess(otus.pp, method = "pca") # seven components are needed
1357
1358 # get sampling locations: 103, 1
1359 otus.group <- c(sample_data(MergedPSGlonPCA)[which(rownames(sample_data(MergedPSGlonPCA)) %in% rownames(otus)), 23])
1360 otus.group <- as.factor(otus.group$Location)
1361
1362 # again, for manuscript, single plot;
1363 bio_pca <- ggbiplot(pca.otus.pp,
1364                   choices = 1:2,
1365                   scale = 1,
1366                   pc.biplot = TRUE,
1367                   ellipse = TRUE,
1368                   varname.size = 4,
1369                   groups = otus.group,
1370
1371                   ) + theme_bw() + theme(legend.title=element_blank(), legend.background = element_rect(color = "gray90",
1372
1373
1374 ## c – Get list of Phyloseq objects, for each Phylum level
1375 PhyseqList <- IsolateTaxa(MergedPSGlon, "Phylum")
1376
1377 ## d – Get Species occurrence and Observation metadata
1378 PhyseqListData <- lapply(PhyseqList, GetSpecObs)
1379
1380 ## e – Calculate empty and full models
1381 PhyseqListModelIsCCA <- lapply(PhyseqListData, ReturnModels, "cca")
1382 # PhyseqListModelIsRDA <- lapply(PhyseqListData, ReturnModels, "rda")
1383 # PhyseqListModelIsCAP <- lapply(PhyseqListData, ReturnModels, "cap")
1384
1385 ## f – Find highest scoring Models for all Constrained Ordination types
1386 BestModCCA <- mapply(ModStep, PhyseqListModelIsCCA, PhyseqListData)
1387 # BestModRDA <- mapply(ModStep, PhyseqListModelIsRDA, PhyseqListData)
1388 # BestModCAP <- mapply(ModStep, PhyseqListModelIsCAP, PhyseqListData)
1389
1390 ## g – plot objects
1391 plot.new()
1392 # par(mfrow = c(1, 4))
1393 mapply(PlotCA, names(BestModCCA), BestModCCA, names(PhyseqList), PhyseqList, site_label = "Location", taxa_level = 3, text_labels = FALSE)
1394 mapply(PlotCA, names(BestModCCA), BestModCCA, names(PhyseqList), PhyseqList, site_label = "Location", taxa_level = 3, text_labels = FALSE)

```

```
1395 mapply(PlotCA, names(BestModCCA), BestModCCA, names(PhyseqList), PhyseqList, site_label = "Location", taxa_level = 3, text_labels = FALSE,
1396
1397 ## h - evaluate model
1398
1399 # test complete model
1400 anova(BestModCCA$All_Taxa)
1401
1402 # test axis
1403 spc <- PhyseqListData$All_Taxa[[1]]
1404 obs <- PhyseqListData$All_Taxa[[2]]
1405 anova(BestModCCA$All_Taxa, by="axis", perm = 1000)
1406
1407 # Type I test
1408 anova(BestModCCA$All_Taxa, by="term", perm = 1000)
1409
1410 # Type III test
1411 anova(BestModCCA$All_Taxa, by="margin", perm = 1000)
1412
1413 # test dropping AICs when adding and removing variables
1414 test_me <- BestModCCA$All_Taxa
1415 vif.cca(test_me)
1416
1417 ## i - generate counts for results section
1418
1419 # sort original (not taxon agglomerated data)
1420 data.frame(tax_table(MergedPS)) [with(data.frame(tax_table(MergedPS)), order(Superphylum, Phylum, Class, Order, Family, Species)), ]
1421
1422 # export as xlsx to wd - see http://www.r-bloggers.com/importexport-data-to-and-from-xlsx-files/
1423 # getwd()
1424 # write.xlsx(data.frame(tax_table(MergedPS)) [with(data.frame(tax_table(MergedPS)), order(Superphylum, Phylum, Class, Order, Family, Species)), ],
1425
1426 ##### 8 - data export
1427
1428 ## a - data export of coordinates for Duanne in KML format
1429
1430 # store sample data in data frame
1431 PCM_md <- data.frame(sample_data(MergedPS))
1432
1433 # define coordinates
1434
```

```
1436 coordinates(PCM_md) <- c("Longitude", "Latitude")
1437
1438 # set projection attributes
1439 proj4string(PCM_md) <- CRS("+proj=longlat +datum=WGS84")
1440
1441 # transform coordinates (not necessary)
1442 # PCM_md_ll <- spTransform(PCM_md, CRS("+proj=longlat +datum=WGS84"))
1443
1444 # set wd for writing shapefiles
1445 setwd("")
1446
1447 # write KML file to current wd
1448 write0GR(PCM_md["Location"], "150602_PCM_regions.kml", layer="region", driver="KML")
1449
1450 # reset to original wd
1451 setwd("")
1452
1453 ## b - data export of sample coordinates for main text figure
1454
1455 # store sample data in data frame
1456 PCM_md <- data.frame(sample_data(MergedPS))
1457
1458 # define coordinates
1459 coordinates(PCM_md) <- c("Longitude", "Latitude")
1460
1461 # set projection attributes
1462 proj4string(PCM_md) <- CRS("+proj=longlat +datum=WGS84")
1463
1464 # set wd for writing shapefiles
1465 setwd("")
1466
1467 # write Esri shapefile file to current wd
1468 write0GR(PCM_md, "150603_PCM_samples", layer="complete_cases", driver="ESRI Shapefile")
1469
1470 # reset to original wd
1471 setwd("")
1472
1473 ##### 9 - counts for writing
1474
1475 ## a - total sample number
```

```

1477 nrow(PCM_md)
1478
1479 ## b – biologic data per Gene
1480
1481 # some samples have incorrect information regarding locus availability
1482 PCM_md[which(PCM_md$Genes == "none"),]$Sample # "2.10.E.PCM" "2.10.C.PCM"
1483
1484 # there is phylotype information available for 18S
1485 otu_table(MergedPS)[,which(colnames(otu_table(MergedPS)) %in% c("2.10.E.PCM", "2.10.C.PCM"))]
1486
1487 # correct metadata – doesn't work yet
1488 PCM_md[ which(PCM_md$Sample == "2.10.E.PCM"), ]$Genes <- "18Sonly"
1489 PCM_md[ which(PCM_md$Sample == "2.10.C.PCM"), ]$Genes <- "18Sonly"
1490
1491 # phylotype information per marker
1492 sum(PCM_md$Genes == "both")
1493 sum(PCM_md$Genes == "18Sonly") # add 2
1494 sum(PCM_md$Genes == "C0Ionly")
1495 PCM_md$Genes
1496
1497
1498

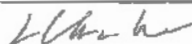
```

D. Molecular tagging of amplicons

Statement of Authorship

Title of Paper	Modular tagging of amplicons using a single PCR for high-throughput sequencing		
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Publication Style		
Publication Details	<p>Clarke, L. J., Czechoński, P., Soubrier, J., Stevens, M. I. & Cooper, A. 2013 Modular tagging of amplicons using a single PCR for high-throughput sequencing. <i>Mol. Ecol. Resour.</i> 14, 1–5. (doi:10.1111/1755-0998.12162)</p> <p>High-throughput sequencing (HTS) of PCR amplicons is becoming the method of choice to sequence one or several targeted loci for phylogenetic and DNA barcoding studies. Although the development of HTS has allowed rapid generation of massive amounts of DNA sequence data, preparing amplicons for HTS remains a rate-limiting step. For example, HTS platforms require platform-specific adapter sequences to be present at the 5' and 3' end of the DNA fragment to be sequenced. In addition, short multiplex identifier (MID) tags are typically added to allow multiple samples to be pooled in a single HTS run. Existing methods to incorporate HTS adapters and MID tags into PCR amplicons are either inefficient, requiring multiple enzymatic reactions and clean-up steps, or costly when applied to multiple samples or loci (fusion primers). We describe a method to amplify a target locus and add HTS adapters and MID tags via a linker sequence using a single PCR. We demonstrate our approach by generating reference sequence data for two mitochondrial loci (COI and 16S) for a diverse suite of insect taxa. Our approach provides a flexible, cost-effective and efficient method to prepare amplicons for HTS.</p>		


Principal Author


Name of Principal Author	Laurence Clarke		
Contribution to the Paper	L.C. led the writing, L.C. and P.C. conducted laboratory work and analysed the results.		
Overall percentage (%)	80%		
Signature		Date	11.6.2015

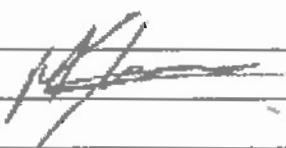
Co-Author Contributions

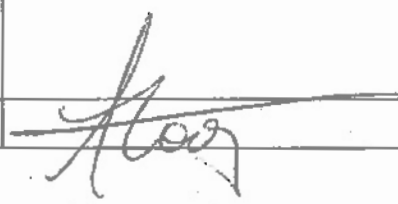
By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all authors contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author (Candidate)	Paul Czechowski		
Contribution to the Paper	L.C. and P.C. conducted laboratory work and analysed the results. P.C contributed to editing the manuscript		
Signature		Date	11.6.2015

Name of Co-Author	Julien Soubrier		
Contribution to the Paper	J.S. designed the bioinformatic analysis and pipeline. J.S. contributed to editing the manuscript.		
Signature		Date	11.06.2015 -

Name of Co-Author	Mark Stevens		
Contribution to the Paper	M.S. helped conceive the experiments, contributed to editing the manuscript.		
Signature		Date	11.6.2015

Name of Co-Author	Alan Cooper		
Contribution to the Paper	A.C. helped conceive the experiments, contributed to editing the manuscript.		
Signature		Date	12.6.2015

Please cut and paste additional co-author panels here as required.

Modular tagging of amplicons using a single PCR for high-throughput sequencing

LAURENCE J. CLARKE,* PAUL CZECHOWSKI,*† JULIEN SOUBRIER,* MARK I. STEVENS†‡ and ALAN COOPER*

*Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005, Australia, †Australian Centre for Evolutionary Biology and Biodiversity, University of Adelaide, Adelaide, SA 5005, Australia, ‡South Australian Museum, GPO Box 234, Adelaide, SA 5000, Australia

Abstract

High-throughput sequencing (HTS) of PCR amplicons is becoming the method of choice to sequence one or several targeted loci for phylogenetic and DNA barcoding studies. Although the development of HTS has allowed rapid generation of massive amounts of DNA sequence data, preparing amplicons for HTS remains a rate-limiting step. For example, HTS platforms require platform-specific adapter sequences to be present at the 5' and 3' end of the DNA fragment to be sequenced. In addition, short multiplex identifier (MID) tags are typically added to allow multiple samples to be pooled in a single HTS run. Existing methods to incorporate HTS adapters and MID tags into PCR amplicons are either inefficient, requiring multiple enzymatic reactions and clean-up steps, or costly when applied to multiple samples or loci (fusion primers). We describe a method to amplify a target locus and add HTS adapters and MID tags via a linker sequence using a single PCR. We demonstrate our approach by generating reference sequence data for two mitochondrial loci (COI and 16S) for a diverse suite of insect taxa. Our approach provides a flexible, cost-effective and efficient method to prepare amplicons for HTS.

Keywords: 16S large ribosomal subunit, amplicon sequencing, cytochrome *c* oxidase subunit I, DNA barcoding, high-throughput sequencing, insect, MoTASP, multiplex identifier

Received 28 May 2013; revision received 8 August 2013; accepted 8 August 2013

Introduction

High-throughput sequencing (HTS) of PCR amplicons (amplicon sequencing) is a common method of targeted sequencing one or several loci in large numbers of samples simultaneously. This approach is regularly employed in targeted resequencing and DNA barcoding projects (e.g. Sørensen *et al.* 2010; Hajibabaei *et al.* 2011; O'Neill *et al.* 2013) to survey biodiversity at multiple levels. Currently available HTS platforms (Illumina, Roche 454, Ion Torrent and SOLiD) require adapter sequences to be present at the 5' and 3' end of the DNA fragment to be sequenced. In addition, short [4–10 base pair (bp)] sample-specific multiplex identifier (MID) tags are typically added to allow multiple samples to be pooled in a single HTS run (Binladen *et al.* 2007; Meyer & Kircher 2010).

Long primer constructs containing HTS adaptors, MID tags and locus-specific primers (fusion primers) can

be used to generate HTS-ready amplicons with a single PCR and are commonly employed for this reason (e.g. Sørensen *et al.* 2010). However, the cost of fusion primers can be prohibitive, particularly when targeting multiple genomic loci, due to separate primers being required for each combination of MID and locus-specific primer (e.g. targeting 10 loci in 20 samples requires 200 primer pairs). An alternative method is to ligate HTS adapters and MID tags to amplicons (Meyer & Kircher 2010; O'Neill *et al.* 2013); however, this requires multiple enzymatic and clean-up steps with concomitant labour costs and risks.

Bybee *et al.* (2011) and de Cárcer *et al.* (2011) demonstrated the potential of using a linker sequence at the 5' end of the locus-specific primer to attach HTS adaptors and MID tags in a modular fashion using a second round of PCR. We develop this approach by amplifying the target locus and attaching adaptors and MID tags via a linker sequence within a single PCR. We use this method to generate COI and 16S rDNA reference sequences for a diverse set of insect taxa to demonstrate the efficiency of the methodology.

Correspondence: Laurence J. Clarke, Fax: +61 8 8313 4364; E-mail: laurence.clarke@adelaide.edu.au

Materials and methods

Insects were collected in the Adelaide Hills (South Australia) using a malaise trap on 2–3 November 2012, and stored in 100% ethanol. *Drymaplaneta communis* (Blattodea) was hand collected in Adelaide on 1 December 2012, and stored in 100% ethanol. Specimens were identified using existing morphological keys to family level. DNA was extracted from 12 specimens using a modified version of the Canadian Centre for DNA Barcoding plate extraction method (Ivanova *et al.* 2006, 2007). A DNA extract from *Ischnura heterosticta* (Odonata) used in a previous study (D. Green, unpublished data) was also included.

We adapted the multiplex-ready PCR method (Hayden *et al.* 2008a) to amplify a target locus and attach adaptors and MID tags in a single PCR prior to HTS. Multiplex-ready PCR (Hayden *et al.* 2008a) uses the principle of M13-tailed primers to add a fluorophore of choice for microsatellite and SNP genotyping in a two-stage PCR. Forward and reverse locus-specific primers are modified to include generic, noncomplementary nucleotide sequences at their 5' ends that act as primer-binding sites in the second stage of PCR. In the first stage of PCR, the locus-specific primers are used to amplify the target loci. In the second stage of PCR, universal primers (*tagF* and *tagR*) tagged with a fluorophore amplify the first-stage products to a detectable level. The involvement of the *tag* primers is restricted to the second stage of the PCR by their lower annealing temperature compared with the locus-specific primers. To add MID tags and HTS adaptors to PCR amplicons (Fig. 1), the fluorophore at the 5' end of the *tagF* primer was replaced by the Ion Torrent Primer P1-key, and the *tagR* primer was modified to include the Primer A-key at the 5' end, followed by a 7-bp MID sequence (Meyer & Kircher 2010).

Forward and reverse linker sequences (corresponding to the *tagF* and *tagR* primer sequences) were added to the 5' end of locus-specific primers targeting the

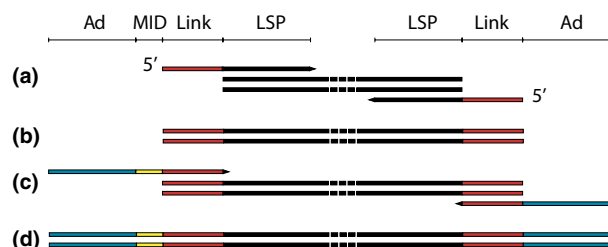


Fig. 1 Schematic representation of modular tagging of amplicons using a single PCR (MoTASP) for high-throughput sequencing (HTS). In the first cycles of PCR, linker sequences (Link) are attached to amplicons via the locus-specific primers (LSP, A and B). In later cycles, HTS adaptors (Ad) and MID tags are attached to amplicons via the linker sequences (C and D).

mitochondrial COI gene and 16S rDNA (Table 1), and loci were PCR amplified using a standardized thermal cycling protocol (Hayden *et al.* 2008b; Appendix S1, Supporting information). PCR was performed in a 12- μ L reaction mixture containing 10 ng genomic DNA, 60 nM of forward and reverse locus-specific primer, 75 nM each of the *tagF* and *tagR* primer constructs, with all other reaction conditions as described in Hayden *et al.* (2008b, Appendix S1, Supporting information). Each insect extract was amplified with a *tagF* primer construct containing a unique MID tag, with the same set of 13 MID tags used for the 16S and COI loci.

PCR products were purified by polyethylene glycol (PEG)/NaCl precipitation with a final concentration of 13% (w/v) PEG, using Sera-Mag carboxylate-modified magnetic speed-beads (Thermo Scientific, Waltham, MA, USA) as the solid phase (DeAngelis *et al.* 1995; Lundin *et al.* 2010). Purified products for each locus were quantified using the Qubit[®] dsDNA HS assay on a Qubit[®] 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and combined in equimolar ratios to create two libraries (COI and 16S). The size distribution and concentration of each library were assessed on an Agilent 2200 TapeStation using High Sensitivity D1K ScreenTape and reagents (Agilent Technologies, Santa Clara, CA, USA). A second PEG precipitation was performed to remove primer artefacts (9% PEG) where necessary. Emulsion PCR and Ion Sphere[™] Particle enrichment were conducted on an Ion OneTouch[™] System (Life Technologies) using the Ion OneTouch[™] 200 Template Kit v2 DL according to the manufacturers' protocol. Each library was sequenced on an Ion Torrent Personal Genome Machine[™] (PGM) Sequencer using the Ion PGM[™] 200 Sequencing Kit and Ion 314[™] semiconductor sequencing chips (Life Technologies).

We developed a customized pipeline to process HTS reads. The script *fastx_barcode_splitter.pl* from the FastX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/) was used to sort reads by MID, using a strict zero mismatches threshold (*-bol -mismatches 0*). The quality of reads assigned to each MID was assessed using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Cutadapt v1.2.1 (Martin 2011) was used to trim linker sequences and locus-specific primers using a maximum error rate of 0.33 (*-e 0.3333*), and to remove short (*-m 25 bp*), long (*-M 220 bp*) and low-quality sequences (*-q 20*), with a total of five passes (*-n 5*). For each locus, trimmed reads were mapped to a reference sequence from the same order or family using Geneious 6.0.3 (Drummond *et al.* 2011), and a consensus sequence was generated by visual inspection of the aligned reads with a minimum of 50-fold coverage. BLAST searches (blastn algorithm, no filter for low complexity or human repeats, search restricted to insect sequences) were used to

Table 1 PCR primers used in this study. Amplicon lengths are given excluding primer sequences. Linker sequences are italics. The position of the multiplex identifier (MID) tag is shown as [N].

Name	Locus	Primer sequence (5'-3')	Length (bp)	Reference
tagF_C1-J-1709	COI	ACGACGTTGTAAAAATTGGWGGWTTTGGAAAYTG	133	Simon <i>et al.</i> (2006)
tagR_C1-N-1843d	COI	CATTAAAGTTCCCATTTAGMWARWGGWGGRTAWACWGTTC		Zhang & Hewitt (1997)
tagF_Ins16S-1F	16S	ACGACGTTGTAAAAATRRGACGAGAGAACCTATA	156	L. J. Clarke and A. Cooper, unpublished
tagR_Ins16S-1Rshort	16S	CATTAAAGTTCCCATTTAACGCTGTTATCCCTAARGTA		L. J. Clarke and A. Cooper, unpublished
ionA_MID_tagR		CCATCTCATCCCTGCGTGTCTCCGACTCAG[NNNNNNN]CATTAAAGTTCCCATTA		
ionP1_tagF		CCTCTCTATGGGCAGTCGGTGATAGAGACACGACGTTGTAAAA		

identify the most similar sequences in the NCBI nucleotide database (Altschul *et al.* 1990).

Results

A total of 511 502 raw reads were obtained for 16S and 329 814 for COI. The number of reads per MID ranged from 9464 to 52 747 (mean \pm SD = $37\,567 \pm 14710$, 23129 did not match one of the 13 MID tags) for 16S and 7232 to 59 673 ($24\,761 \pm 12\,879$, 7918 unmatched) for COI (Table S2, Supporting information). Average Q-values for reads assigned to each MID ranged from 24.2 to 26.8 for 16S (mean \pm SD = 25.6 ± 0.8) and 26.3 to 29.9 for COI (27.9 ± 1.0 , Table S2, Supporting information). After primer trimming, consensus building and BLAST search, the hit with the smallest E-value for each 16S consensus sequence corresponded to the correct insect order in all cases, and to the correct family for 12 of 13 taxa (Table S1, Supporting information). The best BLAST hit for the 13 COI consensus sequences corresponded to the correct order for 10 taxa, with the two hymenopteran sequences retrieving BLAST hits for Lepidoptera. Sanger sequencing using C1-J-1709 and C1-N-1843d (Table 1) or the standard barcoding primers (LCO1490 and HCO2198, Folmer *et al.* 1994) was used to verify that the three COI consensus sequences retrieving BLAST hits to the incorrect order were accurate (100% pairwise identity between consensus and Sanger sequence in each case).

Discussion

Our modular approach to add HTS adapters and MID tags to amplicons in a single PCR provides a simpler and more efficient method than previous protocols (Bybee *et al.* 2011; de Cárcer *et al.* 2011). Incorporating adapters and MID tags in a single, closed-tube reaction reduces the number of clean-up steps required and the potential for contamination. Changing the linker sequence (*tagF* or *tagR*) attached to the locus-specific primer also provides a simple means to change the read direction if desired. The use of a standardized thermal cycling protocol simplifies PCR optimization, such that only locus-specific primer concentration requires optimizing.

We found adapting the multiplex-ready PCR protocol to add HTS adapters and MID tags to amplicons required higher concentrations of locus-specific primers compared with the original application of amplifying microsatellite or SNP loci (Hayden *et al.* 2008a). Hayden *et al.* (2008a,b) recommend testing concentrations of locus-specific primers between 20 and 80 nM to optimize amplification. We have previously found low concentrations (20–40 nM) were most suitable for amplifying microsatellite loci using standard *tagF* and *tagR* primers

(e.g. Clarke *et al.* 2011). When HTS adapters and MID tags are added to the *tagF* and *tagR* primers, low concentrations of locus-specific primers led to large primer artefacts (*ca.* 100 bp) that needed to be removed prior to sequencing. Higher concentrations of locus-specific primer (60–80 nM) led to significant reductions in primer artefacts, presumably due to the presence of more PCR product following the first stage of amplification reducing the extent of primer–primer interactions in the second stage.

These experiments were intended as a proof of concept study; however, by utilizing bioinformatics and HTS capabilities, the number of samples and loci sequenced in a single run could be greatly increased. Although we sequenced the 16S and COI loci on separate Ion 314 chips, multiple loci could be pooled on the same chip using the same MID tags and separated *post hoc* by locus-specific primer sequence. Based on the minimum number of reads for a MID in this study (*ca.* 7200), a sequencing depth of 100-fold for each MID could be obtained by pooling 900 samples, or a combination of samples and loci, for example, 10 loci for 90 samples, on a single 314 chip. An order of magnitude more samples could be pooled using the Ion 316 sequencing chip, with a further 10-fold increase available using the Illumina MiSeq (v2 reagents). Although scaling up our approach to such an extent would require a very large number of unique MIDs, software capable of designing several thousands of unique MIDs is now available (Faircloth & Glenn 2012; Costea *et al.* 2013). Furthermore, the number of primer constructs containing a unique MID could be reduced for large numbers of samples by incorporating MID tags adjacent to the reverse HTS adapter (in this case, the Ion Torrent Primer P1-key). As long as full-length reads are obtained, reads can be separated by the combination of MID tags at the beginning and end of each sequence, greatly increasing the number of unique MID combinations possible. Alternatively, MID tags could be added to the linker-LSP construct and reads sorted based on the combination of two MID tags in a similar fashion.

The modular tagging of amplicons using a single PCR (MoTASP) method can potentially be applied to many loci, with the most critical factor for successful amplification of any given locus being the annealing temperature of the locus-specific primers. We have applied the MoTASP method to several other loci to date, including plant *trnL*, vertebrate 12S rDNA and alternative sites within the COI and 16S loci (L. J. Clarke & P. Czechowski, unpublished data). We have not applied this approach to single-copy nuclear genes as yet, but we expect the method should work as it was originally designed to amplify and fluorescently tag nuclear microsatellites (Hayden *et al.* 2008a). Advances in HTS leading

to increased read lengths (for example, 400-bp kits are now available for the Ion Torrent PGM) will increase the number of loci to which this method can be applied and the amount of data that can be generated, and in turn the taxonomic resolution, for any given locus. In our experience, the most critical factor limiting the application of MoTASP is the annealing temperature of the locus-specific primers. Low annealing temperatures (<50 °C) for the locus-specific primers prevent amplification of the target locus prior to amplification with the linker primers. We have observed improved amplification of some loci by reducing the annealing temperature in the first phase of the thermal cycling protocol (see Appendix S1, Supporting information); however, some loci still failed to amplify. Locus-specific primers with low annealing temperatures could be redesigned (e.g. increased length or GC content) to facilitate amplification with this protocol.

In this study, we have demonstrated a novel method to attach HTS adapters and MID tags to amplicons using a single PCR and used the same set of MID tags to generate reference sequences for two loci commonly used in systematic, barcoding and phylogenetic studies. Modular attachment of HTS adapters and MID tags provides several advantages over the use of fusion primers. A modular approach permits straightforward transfer between HTS platforms by changing the HTS adapter and MID tag primer combined with the locus-specific primer. Furthermore, HTS adapter and MID tag primers can be applied to any locus, allowing transfer between experiments, projects or laboratory groups, representing a substantial cost reduction compared with ordering large numbers of unique fusion primers. MoTASP requires the same number of PCR and clean-up steps as a standard fusion primer approach; hence, laboratory costs are comparable between the two methods. In conclusion, our approach provides a flexible, cost-effective and efficient method to prepare amplicons for high-throughput sequencing.

Acknowledgements

Thanks to Renate Faast and Oliver Wooley for supplying insect specimens, Douglas Green for supplying DNA extracts and John Jennings, Gary Taylor and Remko Leijts for morphological identification. Bastien Llamas helped design the bioinformatic pipeline. We thank the ARC for funding.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, **2**, e197.

- Bybee SM, Bracken-Grissom H, Haynes BD *et al.* (2011) Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, **3**, 1312–1323.
- de Cárcer DA, Denman SE, McSweeney C, Morrison M (2011) Strategy for modular tagged high-throughput amplicon sequencing. *Applied and Environmental Microbiology*, **77**, 6310–6312.
- Clarke LJ, Mackay DA, Whalen MA (2011) Isolation of microsatellites from *Baumea juncea* (Cyperaceae). *Conservation Genetics Resources*, **3**, 113–115.
- Costea PI, Lundeberg J, Akan P (2013) TagGD: fast and accurate software for DNA tag generation and demultiplexing. *PLoS One*, **8**, e57521.
- DeAngelis MM, Wang DG, Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research*, **23**, 4742–4743.
- Drummond AJ, Ashton B, Buxton S *et al.* (2011) *Geneious v5.6.3*. Available from <http://www.geneious.com>.
- Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*, **7**, e42543.
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.
- Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.
- Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008a) Multiplex-Ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics*, **9**, 80.
- Hayden MJ, Nguyen TM, Waterman A, McMichael GL, Chalmers KJ (2008b) Application of multiplex-ready PCR for fluorescence-based SSR genotyping in barley and wheat. *Molecular Breeding*, **21**, 271–281.
- Ivanova NV, deWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.
- Ivanova NV, deWaard JR, Hebert PDN (2007) *CCDB protocols, glass fiber plate DNA extraction*. Available from http://ccdb.ca/docs/CCDB_DNA_Extraction.pdf.
- Lundin S, Stranneheim H, Pettersson E, Klevebring D, Lundeberg J (2010) Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One*, **5**, e10029.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet:journal*, **17**, 10–12.
- Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, doi:10.1101/pdb.prot5448.
- O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, **22**, 111–129.
- Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT (2006) Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Annual Review of Ecology, Evolution and Systematics*, **37**, 545–579.
- Sønstebo JH, Gielly L, Brysting AK *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, **10**, 1009–1018.
- Zhang D-X, Hewitt GM (1997) Assessment of the universality and utility of a set of conserved mitochondrial COI primers in insects. *Insect Molecular Biology*, **6**, 143–150.

L.C. led the writing, L.C. and P.C. conducted laboratory work and analysed the results, J.S. designed the bioinformatic analysis and pipeline, M.S. and A.C. helped conceive the experiments and all authors contributed to editing the manuscript.

Data Accessibility

Contigs used to generate consensus sequences from HTS data (BAM files), consensus sequences (FASTA file) and the bioinformatic pipeline to process HTS reads using the FastX toolkit and Cutadapt are available on Data-Dryad (doi:10.5061/dryad.0f9n0).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Morphological and DNA sequence identification for each specimen in this study.

Table S2 Number of HTS reads assigned and mean Q-values for each MID.

Appendix S1 MoTASP PCR protocol.

Bibliography

- Abouheif, E., R. Zardoya, and A. Meyer (1998). Limitations of metazoan 18S rRNA sequence data: Implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the cambrian explosion. *Journal of Molecular Evolution* 47(4), 394–405.
- Ahn, J. H., B. Y. Kim, J. Song, and H. Y. Weon (2012, December). Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology* 50(6), 1071–1074.
- Altermann, S., S. D. Leavitt, T. Goward, M. P. Nelsen, and H. T. Lumbsch (2014, January). How do you solve a problem like Letharia? A new look at cryptic species in lichen-forming fungi using Bayesian clustering and SNPs from multilocus sequence data. *PLoS ONE* 9(5), e97556.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990, October). Basic local alignment search tool. *Journal of Molecular Biology* 215(3), 403–410.
- Anders, S. and W. Huber (2013). Differential expression of RNA-Seq data at the gene level—the DESeq package.
- Aylagas, E., A. Borja, and N. Rodríguez-Ezpeleta (2014, January). Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). *PLoS ONE* 9(3), e90529.
- Barrett, J., R. Virginia, D. Wall, S. Cary, B. Adams, A. Hacker, and J. Aislabie (2006, December). Co-variation in soil biodiversity and biogeochemistry in northern and southern Victoria Land, Antarctica. *Antarctic Science* 18(04), 535.
- Barrett, J. E., R. A. Virginia, D. H. Wall, and B. J. Adams (2008, August). Decline in a dominant invertebrate species contributes to altered carbon cycling in a low-diversity soil ecosystem. *Global Change Biology* 14(8), 1734–1744.
- Bass, D., A. T. Howe, A. P. Mylnikov, K. Vickerman, E. E. Chao, J. Edwards Smallbone, J. Snell, C. Cabral, and T. Cavalier-Smith (2009, November). Phylogeny and

- classification of Cercomonadida (Protozoa, Cercozoa): Cercomonas, Eocercomonas, Paracercomonas, and Cavernomonas gen. nov. *Protist* 160(4), 483–521.
- Bellemain, E., M. L. Davey, H. v. Kauserud, L. S. Epp, S. Boessenkool, E. Coissac, J. Geml, M. Edwards, E. Willerslev, G. Gussarova, P. Taberlet, and C. Brochmann (2013, May). Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environmental Microbiology* 15(4), 1176–1189.
- Benson, D. a., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2011, January). GenBank. *Nucleic Acids Research* 39(SUPPL. 1), D46–51.
- Berney, C., J. Fahrni, and J. Pawlowski (2004). How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biology* 2(13), 1–13.
- Bik, H. M., D. L. Porazinska, S. Creer, J. G. Caporaso, R. Knight, and W. K. Thomas (2012, May). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution* 27(4), 233–243.
- Bik, H. M., W. Sung, P. De Ley, J. G. Baldwin, J. Sharma, A. Rocha-Olivares, and W. K. Thomas (2012, March). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology* 21(5), 1048–1059.
- Binladen, J., M. T. P. Gilbert, J. P. Bollback, F. Panitz, C. Bendixen, R. Nielsen, and E. Willerslev (2007, January). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2(2), e197.
- Bintanja, R., C. Severijns, R. Haarsma, and W. Hazeleger (2014, June). The future of Antarctica's surface winds simulated by a high-resolution global climate model: II. Drivers of 21st century changes. *Journal of Geophysical Research: Atmospheres* 119(12), 7160–7178.
- Bisby, G. and R. Dennis (1952, December). Notes on british hysteriales. *Transactions of the British Mycological Society* 35(4), 304–307.
- Bockheim, J. G. (1997). Properties and classification of cold desert soils from Antarctica. *Soil Science Society of America Journal* 61(1), 224 – 231.
- Bohmann, K., A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, D. W. Yu, and M. de Bruyn (2014, May). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution* 29(6), 358–367.

- Bokhorst, S., G. K. Phoenix, J. W. Bjerke, T. V. Callaghan, F. Huyer-Brugman, and M. P. Berg (2012, March). Extreme winter warming events more negatively impact small rather than large soil fauna: Shift in community composition explained by traits not taxa. *Global Change Biology* 18(3), 1152–1162.
- Bokulich, N. A., S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso (2013, January). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* 10(1), 57–59.
- Bottos, E. M., J. W. Scarrow, S. D. J. Archer, I. R. McDonald, and S. C. Cary (2014). Bacterial community structures of antarctic soils. In D. A. Cowan (Ed.), *Antarctic Terrestrial Microbiology*, Chapter 2, pp. 9–33. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Bottos, E. M., A. C. Woo, P. Zawar-Reza, S. B. Pointing, and S. C. Cary (2014, January). Airborne Bacterial Populations Above Desert Soils of the McMurdo Dry Valleys, Antarctica. *Microbial Ecology* 67(1), 120–128.
- Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac (2015). OBITools: A Unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources online*(forthcoming), forthcoming.
- Bozdogan, H. (1987, September). Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52(3), 345–370.
- Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson (2013, apr). Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Computational Biology* 9(4), e1003031.
- Budel, B. and C. Colesie (2014). Biological soil crusts. In D. A. Cowan (Ed.), *Antarctic Terrestrial Microbiology*, Chapter 8, pp. 131–161. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Bybee, S. M., H. Bracken-Grissom, B. D. Haynes, R. a. Hermansen, R. L. Byers, M. J. Clement, J. a. Udall, E. R. Wilcox, and K. A. Crandall (2011, January). Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution* 3(1), 1312–1323.
- Campbell, I. B. and G. C. C. Clardidge (Eds.) (1987). *Antarctica: Soils, Weathering Processes and Environment*. Amsterdam: Elsevier.
- Caporaso, J. G., K. Bittinger, F. D. Bushman, T. Z. Desantis, G. L. Andersen, and

- R. Knight (2010, January). PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26(2), 266–267.
- Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight (2010, May). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5), 335–336.
- Caporaso, J. G., C. L. Lauber, W. a. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. a. Gilbert, G. Smith, and R. Knight (2012, August). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* 6(8), 1621–1624.
- Carew, M. E., V. J. Pettigrove, L. Metzeling, and A. a. Hoffmann (2013, January). Environmental monitoring using next generation sequencing: Rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology* 10(1), 45.
- Caruso, T., I. D. Hogg, and R. Bargagli (2010, December). Identifying appropriate sampling and modelling approaches for analysing distributional patterns of Antarctic terrestrial arthropods along the Victoria Land latitudinal gradient. *Antarctic Science* 22(06), 742–748.
- Cavalier-Smith, T. and E. E. Y. Chao (2003, October). Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist* 154(3-4), 341–358.
- CBOL and D. H. Janzen (2009, August). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106(31), 12794–12797.
- Chown, S. L., A. Clarke, C. I. Fraser, S. C. Cary, K. L. Moon, and M. a. McGeoch (2015, June). The changing form of Antarctic biodiversity. *Nature* 522(7557), 431–438.
- Chown, S. L., K. a. Hodgins, P. C. Griffin, J. G. Oakeshott, M. Byrne, and A. a. Hoffmann (2015, January). Biological invasions, climate change and genomics. *Evolutionary Applications* 8(1), 23–46.
- Chown, S. L., A. H. L. Huiskes, N. J. M. Gremmen, J. E. Lee, A. Terauds, K. Crosbie, Y. Frenot, K. a. Hughes, S. Imura, K. Kiefer, M. Lebouvier, B. Raymond, M. Tsujimoto, C. Ware, B. Van de Vijver, and D. M. Bergstrom (2012, March).

- Continent-wide risk assessment for the establishment of nonindigenous species in Antarctica. *Proceedings of the National Academy of Sciences* 109(13), 4938–4943.
- Chown, S. L., J. E. Lee, K. A. Hughes, J. Barnes, P. J. Barrett, D. Bergstrom, P. Convey, D. A. Cowan, K. Crosbie, G. Dyer, Y. Frenot, S. M. Grant, D. Herr, M. C. Kennicutt, M. Lamers, A. Murray, H. P. Possingham, K. Reid, M. J. Riddle, P. G. Ryan, L. Sanson, J. D. Shaw, M. D. Sparrow, C. Summerhayes, A. Terauds, and D. H. Wall (2012). Challenges to the future conservation of the Antarctic. *Science (New York, N.Y.)* 337(July), 158–159.
- Chung, F. H. (1974, December). Quantitative interpretation of X-ray diffraction patterns of mixtures. I. Matrix-flushing method for quantitative multicomponent analysis. *Journal of Applied Crystallography* 7(6), 519–525.
- Clarke, L. J., J. Soubrier, L. S. Weyrich, and A. Cooper (2014, November). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources* 14(6), 1160–1170.
- Cline, J., J. C. Braman, and H. H. Hogrefe (1996, October). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research* 24(18), 3546–3551.
- Coissac, E., T. Riaz, and N. Puillandre (2012, April). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* 21(8), 1834–1847.
- Colesie, C., M. Gommeaux, T. A. Green, and B. Bueddel (2014, April). Biological soil crusts in continental Antarctica: Garwood Valley, southern Victoria Land, and Diamond Hill, Darwin Mountains region. *Antarctic Science* 26(02), 115–123.
- Convey, P. (2003). Soil faunal community response to environmental manipulation on Alexander Island, southern maritime Antarctic. In A. Huiskes, W. Gieskes, J. Rozema, R. Schorno, S. van der Vies, and W. Wolff (Eds.), *Antarctic Biology in a Global Context*, pp. 74–78. Leiden: Backhuys Publishers.
- Convey, P. (2010, August). Terrestrial biodiversity in Antarctica - Recent advances and future challenges. *Polar Science* 4(2), 135–147.
- Convey, P., S. L. Chown, A. Clarke, D. K. A. Barnes, S. Bokhorst, V. Cummings, H. W. Ducklow, F. Frati, T. G. A. Green, S. Gordon, H. J. Griffiths, C. Howard-Williams, A. H. L. Huiskes, J. Laybourn-Parry, W. B. Lyons, A. McMinn, S. A. Morley, L. S. Peck, A. Quesada, S. A. Robinson, S. Schiaparelli, and D. H. Wall (2014, May). The spatial structure of Antarctic biodiversity. *Ecological Monographs* 84(2), 203–244.

- Convey, P., J. A. E. Gibson, C. D. Hillenbrand, D. A. Hodgson, P. J. A. Pugh, J. L. Smellie, and M. I. Stevens (2008, May). Antarctic terrestrial life - challenging the history of the frozen continent? *Biological Reviews* 83(2), 103–117.
- Convey, P. and M. I. Stevens (2007, September). Antarctic biodiversity. *Science (New York, N.Y.)* 317(5846), 1877–1878.
- Convey, P., M. I. Stevens, D. A. Hodgson, J. L. Smellie, C. D. Hillenbrand, D. K. A. Barnes, A. Clarke, P. J. A. Pugh, K. Linse, and S. C. Cary (2009, December). Exploring biological constraints on the glacial history of Antarctica. *Quaternary Science Reviews* 28(27-28), 3035–3048.
- Courtright, E. M., D. H. Wall, and R. A. Virginia (2001, March). Determining habitat suitability for soil invertebrates in an extreme environment: the McMurdo Dry Valleys, Antarctica. *Antarctic Science* 13(01), 9–17.
- Cowan, D. A., J. B. Ramond, T. Makhalanyane, and P. De Maayer (2015, June). Metagenomics of extreme environments. *Current Opinion in Microbiology* 25, 97–102.
- Cowart, D. a., M. Pinheiro, O. Mouchel, M. Maguer, J. Grall, J. Miné, and S. Arnaud-Haond (2015). Metabarcoding is powerful yet still blind: A comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS ONE* 10(2), e0117562.
- Creer, S., V. G. Fonseca, D. L. Porazinska, R. M. Giblin-Davis, W. Sung, D. M. Power, M. Packer, G. R. Carvalho, M. L. Blaxter, P. J. D. Lamshead, and W. K. Thomas (2010, March). Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology* 19(SUPPL. 1), 4–20.
- Cremer, H., D. Gore, N. Hultsch, M. Melles, and B. Wagner (2004, May). The diatom flora and limnology of lakes in the Amery Oasis, East Antarctica. *Polar Biology* 27(9), 513–531.
- Crous, P., R. Shivas, M. Wingfield, B. Summerell, A. Rossman, J. Alves, G. Adams, R. Barreto, A. Bell, M. Coutinho, S. Flory, G. Gates, K. Grice, G. Hardy, N. Kleczewski, L. Lombard, C. Longa, G. Louis-Seize, F. Macedo, D. Mahoney, G. Maresi, P. Martin-Sanchez, L. Marvanová, A. Minnis, L. Morgado, M. Noorde-loos, A. Phillips, W. Quaedvlieg, P. Ryan, C. Saiz-Jimenez, K. Seifert, W. Swart, Y. Tan, J. Tanney, P. Thu, S. Videira, D. Walker, and J. Groenewald (2012, December). Fungal Planet description sheets: 128–153. *Persoonia - Molecular Phylogeny and Evolution of Fungi* 29(1), 146–201.

- Dartnall, H. J. G. (1983). Rotifers of the antarctic and subantarctic. *Hydrobiologia* 104(1), 57–60.
- De Barba, M., C. Miquel, F. Boyer, C. Mercier, D. Rioux, E. Coissac, and P. Taberlet (2014, October). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources* 14(2), 306–323.
- de Cárcer, D. A., S. E. Denman, C. McSweeney, and M. Morrison (2011, September). Strategy for modular tagged high-throughput amplicon sequencing. *Applied and Environmental Microbiology* 77(17), 6310–6312.
- Deagle, B. E., S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet (2014, September). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters* 10(9), 20140562–20140562.
- DeJong, T. M. (1975). A comparison of three diversity indices based on their components of richness and evenness. *Oikos* 26(2), 222–227.
- Delmont, T. O., P. Simonet, and T. M. Vogel (2012, September). Describing microbial communities and performing global comparisons in the 'omic era. *The ISME Journal* 6(9), 1625–1628.
- Delmont, T. O., P. Simonet, and T. M. Vogel (2013, June). Mastering methodological pitfalls for surviving the metagenomic jungle. *BioEssays* 35(8), 744–754.
- Denonfoux, J., N. Parisot, E. Dugat-Bony, C. Biderre-Petit, D. Boucher, D. P. Morgavi, D. L. Paslier, E. Peyretailade, and P. Peyret (2013, April). Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Research* 20(2), 185–196.
- Ding, J., Y. Zhang, Y. Deng, J. Cong, H. Lu, X. Sun, C. Yang, T. Yuan, J. D. Van Nostrand, D. Li, J. Zhou, and Y. Yang (2015, January). Integrated metagenomics and network analysis of soil microbial community of the forest timberline. *Scientific Reports* 5, 7994.
- Dixon, P. (2003, December). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 14(6), 927–930.
- Dolnik, O. V., J. a. C. von Roenn, and S. Bensch (2009, July). *Isospora hypoleucae* sp. n. (Apicomplexa: Eimeriidae), a new coccidian parasite found in the Pied Flycatcher (*Ficedula hypoleuca*). *Parasitology* 136(8), 841–845.
- Dray, S., A. B. Dufour, and D. Chessel (2007). The ade4 package-II: Two-table and K-table methods. *R News* 7(2), 47–52.

- Dreesens, L., C. Lee, and S. Cary (2014, jul). The Distribution and Identity of Edaphic Fungi in the McMurdo Dry Valleys. *Biology* 3(3), 466–483.
- Drummond, A. J., R. D. Newcomb, T. R. Buckley, D. Xie, A. Dopheide, B. C. Potter, J. Heled, H. A. Ross, L. Tooman, S. Grosser, D. Park, N. J. Demetras, M. I. Stevens, J. C. Russell, S. H. Anderson, A. Carter, and N. Nelson (2015, dec). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience* 4(1), 46.
- Edgar, R. C. (2010, October). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19), 2460–2461.
- Edgar, R. C. (2013, October). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10(10), 996–8.
- Efron, B. and C. Stein (1981). The jackknife estimate of variance. *The Annals of Statistics* 9(3), 586–596.
- Egan, S. P., E. Grey, B. Olds, J. Feder, S. Ruggiero, C. E. Tanner, and D. Lodge (2015). Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. *Environmental Science & Technology* 49(7), 4113–4121.
- Elkins, N. Z. and W. G. Whitford (1984, May). The effects of high salt concentration on desert soil microarthropod density and diversity. *The Southwestern Naturalist* 29(2), 239–241.
- Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erseus, V. I. Gusarov, M. E. Eedwards, A. Johnsen, H. K. Stenoeien, K. Hassel, H. Kaus-erhud, N. G. Yoccoz, K. A. Braeathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann (2012, April). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology* 21(8), 1821–1833.
- Ettema, C. H. and D. A. Wardle (2002, April). Spatial soil ecology. *Trends in Ecology and Evolution* 17(4), 177–183.
- Faircloth, B. C. and T. C. Glenn (2012, January). Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS ONE* 7(8), e42543.
- Fell, J. W., G. Scorzetti, L. Connell, and S. Craig (2006, October). Biodiversity of micro-eukaryotes in Antarctic Dry Valley soils with <5% soil moisture. *Soil Biology and Biochemistry* 38(10), 3107–3119.

- Feng, Y.-J., Q.-F. Liu, M.-Y. Chen, D. Liang, and P. Zhang (2015). Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. *Molecular Ecology Resources online*(online), forthcoming.
- Fernández-Mendoza, F., S. Domaschke, M. a. García, P. Jordan, M. P. Martín, and C. Printzen (2011, March). Population structure of mycobionts and photobionts of the widespread lichen *Cetraria aculeata*. *Molecular Ecology* 20(6), 1208–1232.
- Ficetola, G. F., J. Pansu, A. Bonin, E. Coissac, C. Giguët-Covex, M. De Barba, L. Gilly, C. M. Lopes, F. Boyer, F. Pompanon, G. Rayé, and P. Taberlet (2014, October). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* 15(3), 543–556.
- Flemons, P., R. Guralnick, J. Krieger, A. Ranipeta, and D. Neufeld (2007, jan). A web-based GIS tool for exploring the world’s biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). *Ecological Informatics* 2(1), 49–60.
- Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek (1994, October). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3(5), 294–299.
- Freckman, D. W. and R. a. Virginia (1997). Low-diversity antarctic soil nematode communities: Distribution and response to disturbance. *Ecology* 78(2), 363–369.
- Frenot, Y., S. L. Chown, J. Whinam, P. M. Selkirk, P. Convey, M. Skotnicki, and D. M. Bergstrom (2005, February). Biological invasions in the Antarctic: extent, impacts and implications. *Biological Reviews of the Cambridge Philosophical Society* 80(1), 45–72.
- Fretwell, P. T., P. Convey, A. H. Fleming, H. J. Peat, and K. A. Hughes (2011, October). Detecting and mapping vegetation distribution on the Antarctic Peninsula from remote sensing data. *Polar Biology* 34(2), 273–281.
- Gaston, K. J. (2000, May). Global patterns in biodiversity. *Nature* 405(6783), 220–227.
- Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* 15(10), 1451–1455.

- Gilbert, J. a., F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, J. a. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens (2010, January). Meeting report: The terabase metagenomics workshop and the vision of an Earth microbiome project. *Standards in genomic sciences* 3(3), 243–248.
- Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum (2009, February). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27(2), 182–189.
- Gokul, J., A. Valverde, M. Tuffin, S. Cary, and D. Cowan (2013, February). Micro-eukaryotic diversity in hypolithons from Miers Valley, Antarctica. *Biology* 2(1), 331–340.
- Golay, M. (1949). Notes on digital coding. *Proceedings of the Institute of Radio Engineers* 37(6), 657–657.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40(3-4), 237–264.
- Grazulis, S., A. Daskevicius, A. Merkys, D. Chateigner, L. Lutterotti, M. Quiros, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail (2012, January). Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* 40(D1), D420–D427.
- Gutt, J., D. Zurell, T. J. Bracegirdle, W. Cheung, M. S. Clark, P. Convey, B. Danis, B. David, C. De Broyer, G. di Prisco, H. Griffiths, R. Laffont, L. S. Peck, B. Pierrat, M. J. Riddle, T. Saucède, J. Turner, C. Verde, Z. Wang, and V. Grimm (2012, May). Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. *Polar Research* 31(SUPPL.), 1–23.
- Hajibabaei, M., M. A. Smith, D. H. Janzen, J. J. Rodriguez, J. B. Whitfield, and P. D. N. Hebert (2006, July). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes* 6(4), 959–964.
- Hajibabaei, M., J. L. Spall, S. Shokralla, and S. van Konynenburg (2012, December). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology* 12(1), 28.

- Haling, R. E., R. J. Simpson, A. C. McKay, D. Hartley, H. Lambers, K. Ophel-Keller, S. Wiebkin, Herdina, I. T. Riley, and A. E. Richardson (2011, nov). Direct measurement of roots in soil for single and mixed species using a quantitative DNA-based method. *Plant and Soil* 348(1-2), 123–137.
- Hamming, R. W. (1950, apr). Error Detecting and Error Correcting Codes. *Bell System Technical Journal* 29(2), 147–160.
- Hogg, I. D., S. Craig Cary, P. Convey, K. K. Newsham, A. G. O'Donnell, B. J. Adams, J. Aislabie, F. Frati, M. I. Stevens, and D. H. Wall (2006, October). Biotic interactions in antarctic terrestrial ecosystems: are they a factor? *Soil Biology and Biochemistry* 38(10), 3035–3040.
- Howard-Williams, C., I. Hawes, and S. Gordon (2010, December). The environmental basis of ecosystem variability in Antarctica: Research in the Latitudinal Gradient Project. *Antarctic Science* 22(06), 591–602.
- Howard-Williams, C., D. Peterson, W. Lyons, R. Cattaneo-Vietti, and S. Gordon (2006, December). Measuring ecosystem response in a rapidly changing environment: the Latitudinal Gradient Project. *Antarctic Science* 18(4), 465–471.
- Howe, A. T., D. Bass, E. E. Chao, and T. Cavalier-Smith (2011, November). New genera, species, and improved phylogeny of Glissomonadida (Cercozoa). *Protist* 162(5), 710–722.
- Huang, C. Y., H. Kuchel, J. Edwards, S. Hall, B. Parent, P. Eckermann, Herdina, D. M. Hartley, P. Langridge, and A. C. McKay (2013, nov). A DNA-based method for studying root responses to drought in field-grown wheat genotypes. *Scientific Reports* 3, 3194.
- Hughes, K. a. and P. Convey (2012, January). Determining the native/non-native status of newly discovered terrestrial and freshwater species in Antarctica - current knowledge, methodology and management action. *Journal of Environmental Management* 93(1), 52–66.
- Hughes, K. A. and P. Convey (2014, May). Alien invasions in Antarctica - is anyone liable? *Polar Research* 33, 1–13.
- Huson, D. H. and N. Weber (2013, January). Microbial community analysis using MEGAN. *Methods in Enzymology* 531, 465–485.
- Jungblut, A. D., W. F. Vincent, and C. Lovejoy (2012, November). Eukaryotes in arctic and antarctic cyanobacterial mats. *FEMS Microbiology Ecology* 82(2), 416–428.

- Kamenev, E. N., V. a. Glebovitskii, V. P. Kovach, V. S. Semenov, N. L. Alekseev, E. B. Sal'nikova, and V. M. Mikhailov (2009, August). Late Precambrian metamorphic events in Eastern Antarctica (northern Prince Charles Mountains, Radok Lake area, 70°52' S, 67°57' E). *Doklady Earth Sciences* 425(2), 380–383.
- Kanagawa, T. (2003, January). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* 96(4), 317–323.
- Kanda, H. and Y. Mochida (1992). Aquatic mosses found in lakes of the Skarvsnes region, Syowa Station area, Antarctica. In *Proc NIPR Symp on Polar Biol*, Volume 5, pp. 177–179.
- Kappen, L. (2000, May). Some aspects of the great success of lichens in Antarctica. *Antarctic Science* 12(03), 314–324.
- Katana, A., J. Kwiatowski, K. Spalik, B. Zakryś, E. Szalacha, and H. Szymańska (2001, June). Phylogenetic position of *Koliella* (chlorophyta) as inferred from nuclear and chloroplast small subunit rDNA. *Journal of Phycology* 37(3), 443–451.
- Kennedy, A. D. (1994). Simulated climate change - a field manipulation study of polar microarthropod community response to global warming. *Ecography* 17(2), 131–140.
- Kennicutt, M., S. Chown, J. Cassano, D. Liggett, L. Peck, R. Massom, S. Rintoul, J. Storey, D. Vaughan, T. Wilson, I. Allison, J. Ayton, R. Badhe, J. Baeseman, P. Barrett, R. Bell, N. Bertler, S. Bo, A. Brandt, D. Bromwich, S. Cary, M. Clark, P. Convey, E. Costa, D. Cowan, R. Deconto, R. Dunbar, C. Elfring, C. Escutia, J. Francis, H. Fricker, M. Fukuchi, N. Gilbert, J. Gutt, C. Havermans, D. Hik, G. Hosie, C. Jones, Y. Kim, Y. Le Maho, S. Lee, M. Leppe, G. Leitchenkov, X. Li, V. Lipenkov, K. Lochte, J. Lopez-Martinez, C. Luedecke, W. Lyons, S. Marensi, H. Miller, P. Morozova, T. Naish, S. Nayak, R. Ravindra, J. Retamales, C. Ricci, M. Rogan-Finnemore, Y. Ropert-Coudert, A. Samah, L. Sanson, T. Scambos, I. Schloss, K. Shiraishi, M. Siegert, J. Simoes, B. Storey, M. Sparrow, D. Wall, J. Walsh, G. Wilson, J. Winther, J. Xavier, H. Yang, and W. Sutherland (2015, February). A roadmap for Antarctic and Southern Ocean science for the next two decades and beyond. *Antarctic Science* 27(1), 3–18.
- Khan, N., M. Tuffin, W. Stafford, C. Cary, D. C. Lacap, S. B. Pointing, and D. Cowan (2011, November). Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. *Polar Biology* 34(11), 1657–1668.
- Kircher, M., S. Sawyer, and M. Meyer (2012, January). Double indexing overcomes

- inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* 40(1), e3.
- Koch, G. G. (1982). Intraclass correlation coefficient. In *Encyclopedia of Statistical Sciences* (6 ed.), pp. 213–217. John Wiley & Sons, Ltd.
- Koskinen, K., P. Auvinen, K. J. Bjoerkroth, and J. Hultman (2014). Inconsistent denoising and clustering algorithms for amplicon sequence data. *Journal of Computational Biology* 22(00), 1–9.
- Kruskal, W. H. and W. A. Wallis (1952, December). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260), 583–621.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software* 28(5), 1–26.
- Kwong, S., A. Srivathsan, and R. Meier (2012). An update on DNA barcoding: Low species coverage and numerous unidentified sequences. *Cladistics* 28(6), 639–644.
- Lawley, B., S. Ripley, P. Bridge, and P. Convey (2004, October). Molecular analysis of geographic patterns of eukaryotic diversity in antarctic soils. *Applied and Environmental Microbiology* 70(10), 5963–5972.
- Leach, C. M. and M. Tulloch (1972, November). World-wide occurrence of the suspected mycotoxin producing fungus *Drechslera biseptata* with grass seed. *Mycologia* 64(6), 1357–1359.
- Lee, C. K., B. a. Barbier, E. M. Bottos, I. R. McDonald, and S. C. Cary (2012, may). The Inter-Valley Soil Comparative Survey: the ecology of Dry Valley edaphic microbial communities. *The ISME Journal* 6(5), 1046–1057.
- Legendre, P. (2008, March). Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology* 1(1), 3–8.
- Legendre, P. and M. J. Andersson (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69(1), 1–24.
- Lemmon, A. R. and E. M. Lemmon (2012, October). High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology* 61(5), 745–761.
- Lenz, T. L. and S. Becker (2008, December). Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. *Gene* 427(1-2), 117–123.

- Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, J. T. Boehm, and R. J. Machida (2013, June). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology* 10(1), 34.
- Limtong, S., R. Kaewwichian, W. Yongmanitchai, and H. Kawasaki (2014, June). Diversity of culturable yeasts in phylloplane of sugarcane in Thailand and their capability to produce indole-3-acetic acid. *World Journal of Microbiology and Biotechnology* 30(6), 1785–1796.
- Lindgreen, S. (2012, January). AdapterRemoval: Easy cleaning of next generation sequencing reads. *BMC Research Notes* 5(1), 337.
- Liu, S., Y. Li, J. Lu, X. Su, M. Tang, R. Zhang, L. Zhou, C. Zhou, Q. Yang, Y. Ji, D. W. Yu, and X. Zhou (2013, December). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution* 4(12), 1142–1150.
- Lohse, M., A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt, and B. Usadel (2012, June). RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* 40(W1), W622–W627.
- López-Bueno, A., J. Tamames, D. Velázquez, A. Moya, A. Quesada, and A. Alcamí (2009, November). High diversity of the viral community from an Antarctic lake. *Science (New York, N.Y.)* 326(5954), 858–861.
- Lozupone, C. and R. Knight (2005, December). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71(12), 8228–8235.
- Lozupone, C., M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight (2011). UniFrac: An effective distance metric for microbial community comparison. *The ISME Journal* 5(2), 169–172.
- Machida, R. J. and N. Knowlton (2012, January). PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS ONE* 7(9), e46180.
- Magalhaes, C., M. I. Stevens, S. C. Cary, B. A. Ball, B. C. Storey, D. H. Wall, R. Tuerk, and U. Ruprecht (2012). At limits of life: Multidisciplinary insights reveal environmental constraints on biotic diversity in continental Antarctica. *PLoS ONE* 7(9), 1–10.
- Makhalanyane, T. P., A. Valverde, N.-K. r. Birkeland, S. C. Cary, I. M. Tuffin, and

- D. a. Cowan (2013, June). Evidence for successional development in Antarctic hypolithic bacterial communities. *The ISME Journal* 7(11), 2080–90.
- Marchant, D. R. and J. W. Head (2007, December). Antarctic dry valleys: microclimate zonation, variable geomorphic processes, and implications for assessing climate change on Mars. *Icarus* 192(1), 187–222.
- McGaughran, A., M. I. Stevens, I. D. Hogg, and A. Carapelli (2011, April). Extreme glacial legacies: A synthesis of the Antarctic springtail phylogeographic record. *Insects* 2(2), 62–82.
- McGeoch, M. a., J. D. Shaw, A. Terauds, J. E. Lee, and S. L. Chown (2015, May). Monitoring biological invasion across the broader Antarctic: A baseline and indicator framework. *Global Environmental Change* 32, 108–125.
- McMurdie, P. J. and S. Holmes (2013, January). Phyloseq: An R Package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8(4), e61217.
- McMurdie, P. J. and S. Holmes (2014, April). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology* 10(4), e1003531.
- Medlin, L., H. J. Elwood, S. Stickel, and M. L. Sogin (1988, November). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* 71(2), 491–499.
- Metzker, M. L. (2010, January). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11(1), 31–46.
- Meyer, M., U. Stenzel, and M. Hofreiter (2008, January). Parallel tagged sequencing on the 454 platform. *Nature Protocols* 3(2), 267–278.
- Moberg, R. and T. H. Nash (1999, January). The genus *Heterodermia* in the Sonoran Desert area. *The Bryologist* 102(1), 1–14.
- Moritz, C., T. E. Dowling, and W. M. Brown (1987, April). Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* 18(1), 269–292.
- Nagao, M., K. Matsui, and M. Uemura (2008, June). *Klebsormidium flaccidum*, a charophycean green alga, exhibits cold acclimation that is closely associated with compatible solute accumulation and ultrastructural changes. *Plant, Cell and Environment* 31(6), 872–885.
- Nakai, R., T. Abe, T. Baba, S. Imura, H. Kagoshima, H. Kanda, Y. Kohara, A. Koi, H. Niki, K. Yanagihara, and T. Naganuma (2012, May). Eukaryotic phylotypes in

- aquatic moss pillars inhabiting a freshwater lake in East Antarctica, based on 18S rRNA gene analysis. *Polar Biology* 35(10), 1495–1504.
- Nash III, T. H. (1972, January). Simplification of the Blue Mountain lichen communities near a zinc factory. *The Bryologist* 75(3), 315–324.
- Navarro-Noya, Y. E., A. Jiménez-Aguilar, C. Valenzuela-Encinas, R. J. Alcántara-Hernández, V. M. Ruíz-Valdiviezo, A. Ponce-Mendoza, M. Luna-Guido, R. Marsch, and L. Dendooven (2013, February). Bacterial communities in soil under moss and lichen-moss crusts. *Geomicrobiology Journal* 31(October), 152–160.
- Niederberger, T. D., J. a. Sohm, T. E. Gunderson, A. E. Parker, J. Tirindelli, D. G. Capone, E. J. Carpenter, and S. C. Cary (2015, jan). Microbial community composition of transiently wetted Antarctic Dry Valley soils. *Frontiers in Microbiology* 6(January), 1–12.
- Nielsen, U. N. and D. H. Wall (2013, March). The future of soil invertebrate communities in polar regions: different climate change responses in the Arctic and Antarctic?
- Nielsen, U. N., D. H. Wall, B. J. Adams, and R. a. Virginia (2011, April). Antarctic nematode communities: observed and predicted responses to climate change. *Polar Biology* 34(11), 1701–1711.
- Nkem, J., R. Virginia, J. Barrett, D. Wall, and G. Li (2006, July). Salt tolerance and survival thresholds for two species of Antarctic soil nematodes. *Polar Biology* 29(8), 643–651.
- Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner (2015). Vegan: Community ecology package.
- O'Neill, E. M., R. Schwartz, C. T. Bullock, J. S. Williams, H. B. Shaffer, X. Aguilar-Miguel, G. Parra-Olea, and D. W. Weisrock (2013, October). Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* 22(1), 111–129.
- Ophel-Keller, K., A. McKay, D. Hartley, Herdina, and J. Curran (2008). Development of a routine DNA-based testing service for soilborne diseases in Australia. *Australasian Plant Pathology* 37(3), 243–253.
- Orgiazzi, A., V. Bianciotto, P. Bonfante, S. Daghino, S. Ghignone, A. Lazzari, E. Lumini, A. Mello, C. Napoli, S. Perotto, A. Vizzini, S. Bagella, C. Murat, and

- M. Girlanda (2013, February). 454 pyrosequencing analysis of fungal assemblages from geographically distant, disparate soils reveals spatial patterning and a core mycobiome. *Diversity* 5(1), 73–98.
- Otto, J. (1993, March). A new species of *Microcaeculus* from Australia (Acarina: Caeculidae), with notes on its biology and behavior. *International Journal of Acarology* 19(1), 3–13.
- Otto, J. and K. Wilson (2001). Assessment of the usefulness of ribosomal 18S and mitochondrial COI sequences in Prostigmata phylogeny. In *Acarology: Proceedings of the 10th International Congress.*, Melbourne. CSIRO PUBLISHING.
- Palmer, M. W. (1993, December). Putting things in even better order: The advantages of Canonical Correspondence Analysis. *Ecology* 74(8), 2215–2230.
- Pankhurst, C. E., K. Ophel-Keller, B. M. Doube, and V. V. S. R. Gupta (1996). Biodiversity of soil microbial communities in agricultural systems. *Biodiversity and Conservation* 5(2), 197–209.
- Parchert, K. J., M. N. Spilde, A. Porras-Alfaro, A. M. Nyberg, and D. E. Northup (2012, October). Fungal communities associated with rock varnish in Black Canyon, New Mexico: casual inhabitants or essential partners? *Geomicrobiology Journal* 29(8), 752–766.
- Parfrey, L. W., W. a. Walters, C. L. Lauber, J. C. Clemente, D. Berg-Lyons, C. Teiling, C. Kodira, M. Mohiuddin, J. Brunelle, M. Driscoll, N. Fierer, J. a. Gilbert, and R. Knight (2014, jun). Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Frontiers in Microbiology* 5(JUN), 1–13.
- Peat, H. J., A. Clarke, and P. Convey (2007, August). Diversity and biogeography of the antarctic flora. *Journal of Biogeography* 34(1), 132–146.
- Pedersen, M. W., S. Overballe-Petersen, L. Ermini, C. D. Sarkissian, J. Haile, M. Hellstrom, J. Spens, P. F. Thomsen, K. Bohmann, E. Cappellini, I. B. Schnell, N. A. Wales, C. Caroe, P. F. Campos, A. M. Z. Schmidt, M. T. P. Gilbert, A. J. Hansen, L. Orlando, and E. Willerslev (2014, dec). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1660), 20130383–20130383.
- Pérez, G., B. Slippers, M. J. Wingfield, B. D. Wingfield, A. J. Carnegie, and T. I. Burgess (2012, September). Cryptic species, native populations and biological invasions by a eucalypt forest pathogen. *Molecular Ecology* 21(18), 4452–4471.

- Pointing, S. B. and J. Belnap (2012, August). Microbial colonization and controls in dryland systems. *Nature Reviews Microbiology* 10(9), 654–654.
- Pointing, S. B., Y. Chan, D. C. Lacap, M. C. Y. Lau, J. A. Jurgens, and R. L. Farrell (2009, nov). Highly specialized microbial diversity in hyper-arid polar desert. *Proceedings of the National Academy of Sciences of the United States of America* 106(47), 12009–19964.
- Powers, L. E., M. Ho, D. W. Freckman, and R. A. Virginia (1998, May). Distribution, Community Structure, and Microhabitats of Soil Invertebrates along an Elevational Gradient in Taylor Valley, Antarctica. *Arctic and Alpine Research* 30(2), 133–141.
- Price, M. N., P. S. Dehal, and A. P. Arkin (2009, July). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* 26(7), 1641–1650.
- Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Gloeckner (2007, January). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35(21), 7188–7196.
- Pugh, P. (1993, April). A synonymic catalogue of the Acari from Antarctica, the sub-Antarctic Islands and the Southern Ocean. *Journal of Natural History* 27(2), 323–421.
- R Development Core Team (2011). R: A language and environment for statistical computing.
- Rao, S., Y. Chan, D. C. Lacap, K. D. Hyde, S. B. Pointing, and R. L. Farrell (2012, apr). Low-diversity fungal assemblage in an Antarctic Dry Valleys soil. *Polar Biology* 35(4), 567–574.
- Ratnasingham, S. and P. D. N. Hebert (2007, May). BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes* 7(3), 355–364.
- Rayment, G. E. and D. J. Lyons (2011). *Soil Chemical Methods - Australasia*. Collingwood: CSIRO publishing.
- Reid, M. K. and K. L. Spencer (2009). Use of principal components analysis (PCA) on estuarine sediment datasets: The effect of data pre-treatment. *Environmental Pollution* 157(8-9), 2275–2281.
- Riaz, T., W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, and E. Coissac (2011, November). EcoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research* 39(21), e145.

- Robertson, C. E., J. K. Harris, B. D. Wagner, D. Granger, K. Browne, B. Tatem, L. M. Feazel, K. Park, N. R. Pace, and D. N. Frank (2013, December). Explicit: Graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics* 29(23), 3100–3101.
- Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), 139–140.
- Roche (2009a). TCB No. 005-2009 - using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium chemistry - extended MID Set.
- Roche (2009b). TCB No. 013-2009 - amplicon fusion primer design guidelines for GS FLX Titanium series Lib-A chemistry.
- Roesch, L. F. W., R. R. Fulthorpe, A. B. Pereira, C. K. Pereira, L. N. Lemos, A. D. Barbosa, A. K. A. Suleiman, A. L. Gerber, M. G. Pereira, A. Loss, and E. M. da Costa (2012, October). Soil bacterial community abundance and diversity in ice-free areas of Keller Peninsula, Antarctica. *Applied Soil Ecology* 61, 7–15.
- Rogers, A. D. (2007, December). Evolution and biodiversity of antarctic organisms: a molecular perspective. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 362(1488), 2191–2214.
- Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich (1988, January). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)* 239(4839), 487–491.
- Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician* 41(2), 169.
- Salter, S., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. Loman, and A. W. Walker (2014, January). Reagent contamination can critically impact sequence-based microbiome analyses. Technical Report 87.
- Sansom, J. (1989, oct). Antarctic Surface Temperature Time Series. *Journal of Climate* 2(10), 1164–1172.
- Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. a. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber (2009, December). Introducing mothur: Open-source, platform-independent, community-supported software for

- describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23), 7537–7541.
- Shannon, C. (1948, July). A mathematical theory of communication. *Bell System Technology Journal* 27(3), 379:423, 623–656.
- Shaw, J. D., A. Terauds, M. J. Riddle, H. P. Possingham, and S. L. Chown (2014). Antarctica’s Protected Areas Are Inadequate, Unrepresentative, and at Risk. *PLoS Biology* 12(6), 1–5.
- Shokralla, S., G. Singer, and M. Hajibabaei (2010, mar). Direct PCR amplification and sequencing of specimens’ DNA from preservative ethanol. *BioTechniques* 48(3), 233–234.
- Simpson, E. H. (1949, April). Measurement of diversity. *Nature* 163(4148), 688–688.
- Sinclair, B. J. (2001, May). On the distribution of terrestrial invertebrates at Cape Bird, Ross Island, Antarctica. *Polar Biology* 24(6), 394–400.
- Sinclair, B. J., M. B. Scott, C. J. Klok, J. S. Terblanche, D. J. Marshall, B. Reyers, and S. L. Chown (2006). Determinants of terrestrial arthropod community composition at Cape Hallett, Antarctica. *Antarctic Science* 18(3), 303–312.
- Skotnicki, M. L., P. M. Selkirk, and S. D. Boger (2012, April). New records of three moss species (*Ptychostomum pseudotriquetrum*, *Schistidium antarctici*, and *Coscinodon lawianus*) from the southern Prince Charles Mountains, Mac Robertson Land, Antarctica. *Polar Record* 48(04), 394–400.
- Smith, D. P. and K. G. Peay (2014, January). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE* 9(2), e90234.
- Sohlenius, B. and S. Bostroem (2008, June). Species diversity and random distribution of microfauna in extremely isolated habitable patches on Antarctic nunataks. *Polar Biology* 31(7), 817–825.
- Sohlenius, B., S. Bostroem, and A. Hirschfelder (1996, March). Distribution patterns of microfauna (nematodes, rotifers and tardigrades) on nunataks in Dronning Maud Land, East Antarctica. *Polar Biology* 16(3), 191–200.
- Southwell, C., J. McKinlay, M. Low, D. Wilson, K. Newbery, J. L. Lieser, and L. Emmerson (2013, March). New methods and technologies for regional-scale abundance estimation of land-breeding marine animals: application to Adélie penguin populations in East Antarctica. *Polar Biology* 36(6), 843–856.
- Stenøien, H. K. (2008, March). Slow molecular evolution in 18S rDNA, rbcL

- and nad5 genes of mosses compared with higher plants. *Journal of Evolutionary Biology* 21(2), 566–571.
- Stevens, M. I. and I. D. Hogg (2003, July). Long-term isolation and recent range expansion from glacial refugia revealed for the endemic springtail *Gomphiocephalus hodgsoni* from Victoria Land, Antarctica. *Molecular Ecology* 12(9), 2357–2369.
- Stiller, M., M. Knapp, U. Stenzel, M. Hofreiter, and M. Meyer (2009, October). Direct multiplex sequencing (DMPS) - a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research* 19(10), 1843–1848.
- Sutherland, W. J., W. M. Adams, R. B. Aronson, R. Aveling, T. M. Blackburn, S. Broad, G. Ceballos, I. M. Côté, R. M. Cowling, G. a. B. Da Fonseca, E. Dinerstein, P. J. Ferraro, E. Fleishman, C. Gascon, M. Hunter, J. Hutton, P. Kareiva, A. Kuria, D. W. MacDonald, K. MacKinnon, F. J. Madgwick, M. B. Mascia, J. McNeely, E. J. Milner-Gulland, S. Moon, C. G. Morley, S. Nelson, D. Osborn, M. Pai, E. C. M. Parsons, L. S. Peck, H. Possingham, S. V. Prior, a. S. Pullin, M. R. W. Rands, J. Ranganathan, K. H. Redford, J. P. Rodriguez, F. Seymour, J. Sobel, N. S. Sodhi, A. Stott, K. Vance-Borland, and a. R. Watkinson (2009, June). One hundred questions of importance to the conservation of global biological diversity. *Conservation Biology* 23(3), 557–567.
- Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev (2012, April). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21(8), 2045–2050.
- Taberlet, P., S. M. Prud'Homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, J. C. Clément, C. Melodelima, F. Pompanon, and E. Coissac (2012, February). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology* 21(8), 1816–1820.
- Tang, C. Q., F. Leasi, U. Obertegger, A. Kieneke, T. G. Barraclough, and D. Fontaneto (2012, October). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences* 109(40), 16208–16212.
- Tedrow, J. C. F. (1966). Polar Desert Soils. *Soil Science Society of America Journal* 30(3), 381–387.
- Teixeira, L. C. R. S., R. S. Peixoto, J. C. Cury, W. J. Sul, V. H. Pellizari, J. Tiedje,

- and A. S. Rosado (2010, August). Bacterial diversity in rhizosphere soil from antarctic vascular plants of Admiralty Bay, maritime Antarctica. *The ISME Journal* 4(8), 989–1001.
- Ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis.
- Terauds, A., S. L. Chown, F. Morgan, H. J. Peat, D. J. Watts, H. Keys, P. Convey, and D. M. Bergstrom (2012, July). Conservation biogeography of the Antarctic. *Diversity and Distributions* 18(7), 726–741.
- Tikhonenkov, D. (2013, June). Species diversity and changes of communities of heterotrophic flagellates (protista) in response to glacial melt in King George Island, the South Shetland Islands, Antarctica. *Antarctic Science* 26(02), 133–144.
- Tingey, R. J. (1974, October). Australian geological mapping in the Prince Charles Mountains, 1968–73. *Polar Record* 17(107), 150.
- Tréguier, A., J. M. Paillisson, T. Dejean, A. Valentini, M. a. Schlaepfer, and J. M. Roussel (2014). Environmental DNA surveillance for invertebrate species: Advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds. *Journal of Applied Ecology* 51(4), 871–879.
- Treonis, A. M., D. H. Wall, and R. A. Virginia (1999, October). Invertebrate Biodiversity in Antarctic Dry Valley Soils and Sediments. *Ecosystems* 2(6), 482–492.
- Tukey, J. W. (1949, June). Comparing individual means in the analysis of variance. *Biometrics* 5(2), 99–114.
- Turner, J., N. E. Barrand, T. J. Bracegirdle, P. Convey, D. a. Hodgson, M. Jarvis, A. Jenkins, G. Marshall, M. P. Meredith, H. Roscoe, J. Shanklin, J. French, H. Goosse, M. Guglielmin, J. Gutt, S. Jacobs, M. C. Kennicutt, V. Masson-Delmotte, P. Mayewski, F. Navarro, S. Robinson, T. Scambos, M. Sparrow, C. Summerhayes, K. Speer, and A. Klepikov (2013, April). Antarctic climate change and the environment: an update. *Polar Record* 50(3), 1–23.
- Vallone, P. M. and J. M. Butler (2004, August). AutoDimer: A screening tool for primer-dimer and hairpin structures. *BioTechniques* 37(2), 226–231.
- Van Dijk, E. L., H. Auger, Y. Jaszczyszyn, and C. Thermes (2014, sep). Ten years of next-generation sequencing technology. *Trends in Genetics* 30(9), 418–426.
- Van Rossum, G. and F. L. Drake Jr. (1995). *Python tutorial*. Amsterdam: Centrum voor Wiskunde en Informatica.

- Velasco-Castrillón, A., J. a. E. Gibson, and M. I. Stevens (2014, July). A review of current Antarctic limno-terrestrial microfauna. *Polar Biology* 37(10), 1517–1531.
- Velasco-Castrillón, A., T. J. Page, J. A. E. Gibson, and M. I. Stevens (2014, July). Surprisingly high levels of biodiversity and endemism amongst Antarctic rotifers uncovered with mitochondrial DNA. *Biodiversity* 15(2-3), 130–142.
- Velasco-Castrillón, A., M. B. Schultz, F. Colombo, J. a. E. Gibson, K. a. Davies, A. D. Austin, and M. I. Stevens (2014, January). Distribution and diversity of soil microfauna from East Antarctica: assessing the link between biotic and abiotic factors. *PLoS ONE* 9(1), e87529.
- Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. a. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith (2004, April). Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)* 304(5667), 66–74.
- Voglmayr, H. and W. M. Jaklitsch (2011, January). Molecular data reveal high host specificity in the phylogenetically isolated genus *Massaria* (Ascomycota, Massariaceae). *Fungal Diversity* 46(1), 133–170.
- Wagner, B., H. Cremer, N. Hultsch, D. B. Gore, and M. Melles (2004, November). Late Pleistocene and Holocene history of Lake Terrasovoje, Amery Oasis, East Antarctica, and its climatic and environmental implications. *Journal of Paleolimnology* 32(4), 321–339.
- Wall, D. H. (2012, February). Global Change in a Low Diversity Terrestrial Ecosystem: the McMurdo Dry Valleys. In A. D. Rogers, N. M. Johnston, E. J. Murphy, and A. Clarke (Eds.), *Antarctic Ecosystems: An Extreme Environment in a Changing World*, Chapter 2. Chichester, UK: John Wiley & Sons, Ltd.
- Wall, D. H. and R. a. Virginia (1999, October). Controls on soil biodiversity: insights from extreme environments. *Applied Soil Ecology* 13(2), 137–150.
- White, D. (2007). *Cenozoic glacial history and landscape evolution of Mac. Robertson Land and the Lambert Glacier-Amery ice shelf system, East Antarctica*. Phd thesis, Macquarie University.
- White, D. A., D. Fink, and D. B. Gore (2011, December). Cosmogenic nuclide evidence for enhanced sensitivity of an East Antarctic ice stream to change during the last deglaciation. *Geology* 39(1), 23–26.

- White, D. A. and W. D. Hermichen (2007). Glacial and periglacial history of the southern Prince Charles Mountains, East Antarctica. *Terra Antartica* 14(1), 5–12.
- Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software* 21(12), 1–20.
- Wickham, H. (2009). *Ggplot2 - elegant graphics for data analysis*, Volume 1. Dordrecht Heidelberg London New York: Springer.
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software* 40(1), 1–29.
- Wiemers, M. and K. Fiedler (2007). Does the DNA barcoding gap exist? - A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* 4(September 2004), 8.
- Wilcox, L., P. Fuerst, and G. Floyd (1993). Phylogenetic relationships of four charophycean green algae inferred from complete nuclear-encoded small subunit rRNA gene sequences. *American Journal of Botany* 80(9), 1028–1033.
- Wish, M. and J. D. Carroll (1982). Multidimensional scaling and its applications. In *Handbook of Statistics*, Volume 2, Chapter 14, pp. 317–345. North-Holland: Elsevier B.V.
- Wu, T., E. Ayres, R. D. Bardgett, D. H. Wall, and J. R. Garey (2011, October). Molecular study of worldwide distribution and diversity of soil animals. *Proceedings of the National Academy of Sciences* 108(43), 17720–17725.
- Wu, T., E. Ayres, G. Li, R. D. Bardgett, D. H. Wall, and J. R. Garey (2009). Molecular profiling of soil animal diversity in natural ecosystems: Incongruence of molecular and morphological results. *Soil Biology and Biochemistry* 41(4), 849–857.
- Yu, D. W., Y. Ji, B. C. Emerson, X. Wang, C. Ye, C. Yang, and Z. Ding (2012, August). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* 3(4), 613–623.
- Zawierucha, K., J. Smykla, Ł. Michalczyk, B. Gołdyn, and Ł. Kaczmarek (2015, February). Distribution and diversity of Tardigrada along altitudinal gradients in the Hornsund, Spitsbergen (Arctic). *Polar Research* 34, 24168.
- Zhan, A., S. A. Bailey, D. D. Heath, and H. J. Macisaac (2014). Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources* 14(5), 1049–1059.
- Zhan, A., W. Xiong, S. He, and H. J. MacIsaac (2014, May). Influence of artifact

removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS ONE* 9(5), e96928.

Zhang, D. X. and G. M. Hewitt (1997, May). Assessment of the universality and utility of a set of conserved mitochondrial COI primers in insects. *Insect Molecular Biology* 6(2), 143–150.