



Bioprospecting and Functional Analysis of Neglected Environments

Vogt, Josef Korbinian; Sicheritz-Pontén, Thomas

Publication date:
2013

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):

Vogt, J. K., & Sicheritz-Pontén, T. (2013). Bioprospecting and Functional Analysis of Neglected Environments. Technical University of Denmark (DTU).

DTU Library

Technical Information Center of Denmark

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Bioprospecting and Functional Analysis of Neglected Environments

Josef Korbinian Vogt

30th November, 2013

CENTERFO
R BIOLOGI
CAL SEQU
ENCE ANA
LYSIS **CBS**

Preface

This thesis was prepared at the Center for Biological Sequence Analysis, Department of Systems Biology, at the Technical University of Denmark in partial fulfillment of acquiring the PhD degree. The PhD was funded by the NOVENIA (Enzymes of Industrial Relevance) project and DTU.

All the work was carried out at the Center for Biological Sequence Analysis under supervision of Professor Thomas Sicheritz-Pontén.

Lyngby, November 2013

A handwritten signature in black ink, reading "Josef K Vogt". The signature is written in a cursive style with a large, stylized 'V' at the end.

Josef Korbinian Vogt

Contents

Preface	iii
Contents	v
Abstract	vii
Dansk resumé	viii
Acknowledgements	x
Papers included in the thesis	xi
Papers not included in the thesis	xii
I Introduction	1
1 General Introduction	3
2 Next Generation Sequencing	7
2.1 Metagenomics	10
2.2 Whole Transcriptome Shotgun Sequencing	13
3 Governance of Environmental Samples	15
3.1 Bioprospecting as a research strategy	15
3.2 Biopiracy	16
3.3 Convention on Biological Diversity	17
4 Proteolytic enzymes	19
4.1 Classification	19
4.2 Proteolytic enzymes in the industry	21
II Methods	23
5 From Sequencing Reads to Sequence Assembly	25
5.1 Pre-Processing of Sequencing reads	25
5.2 <i>De novo</i> Assembly	27
5.3 Assembly assessment	30

6	Gene finding and <i>ab initio</i> prediction	31
6.1	Gene finding in metagenomic datasets	32
6.2	Gene finding in eukaryotes	33
7	From Sequence to Function and Taxonomy	35
7.1	Functional Annotation of Coding Regions	35
7.2	Taxonomic Annotation of Metagenomes	38
7.3	Manuscript I	40
8	Analysis of polar marine environments	49
8.1	From Genes to Abundance matrix	49
8.2	Functional analysis – Finding the needle in the haystack . . .	51
8.3	Identifying proteolytic enzymes in the polar marine environment	52
III	Manuscripts	53
9	Manuscript II	55
9.1	The polar marine environment	55
9.2	Comparative functional analysis of Arctic marine metagenomes reveals strategies for deep sea persistence . . .	58
10	Manuscript III	79
10.1	Proteolytic enzymes of the polar marine environment - Patent application	79
10.2	Exploiting the polar marine environment for bioprospecting: novel protease discovery	85
11	Manuscript IV	99
11.1	Carnivorous plants - the Venus flytrap	99
11.2	Transcriptome and genome analyses of the Venus flytrap (<i>D. muscipula</i>)	101
IV	Epilogue	113
12	Summary and Perspectives	115
12.1	Summary	115
	Bibliography	119
	Appendices	131

Abstract

Advances in Next Generation Sequencing technologies made it possible to sequence DNA extracted from environments and organisms at a reasonable cost allowing research fields such as metagenomics and whole transcriptome sequencing (RNA-seq) to be established. These techniques allow the study of functional relationships in single organisms and environments. The sequencing data can also be mined for novel compounds and enzymes. The process of exploiting biological resources for commercial use is known as bioprospecting.

This PhD thesis describes the concept of bioprospecting in the post genomic era (Chapter 1) and introduces the research fields of metagenomics and RNA-seq (Chapter 2) as concepts to access and analyze biological resources. When attempting to discover and commercialize such biological resources, legal obligations have to be met, which is generally governed by the Convention on Biological Diversity (explained in Chapter 3). Proteolytic enzymes – described in Chapter 4 – are the target for bioprospecting due to their high market value. Section II describes methods used for the analysis of metagenomic and RNA-seq datasets, including Manuscript I, which includes the taxonomic annotation of a late Pleistocene horse metagenome and the functional annotation of the donkey genome. The functional analysis and the identification of novel proteolytic enzymes in the polar marine environment and the full transcriptome analysis of the carnivorous plant *Dionaea muscipula* is also presented.

The polar seas are a unique, extreme habitat with constant low temperatures and no light penetration in the deep. Water samples at varying depth (40 m – 4,300 m) were collected during the Galathea III and LOMROG II polar expeditions. The sample DNA was extracted and sequenced. Comparative functional analysis of arctic marine metagenomes reveals bacterial strategies for deep sea persistence (Manuscript II). Furthermore, this extreme environment is a fertile ground to mine for novel proteolytic enzymes. Manuscript III presents a bioinformatics approach to identify sequences for potential commercialization.

Carnivory is a rare trait in the plant kingdom, and only few species are able to trap and digest prey. The sequencing, assembly and functional annotation of a normalized transcriptome of the most famous carnivorous plant, the Venus flytrap (*Dionaea muscipula*), is presented in Manuscript IV.

Chapter 12 summarizes the thesis and includes final remarks on the future perspectives on the presented research. In summary, this thesis demonstrates how biological resources can be exploited for commercial use. Furthermore, the findings give a better understanding of the microbial community's persistence in the deep sea. Lastly, the transcriptome data of the Venus flytrap provide a public resource for unveiling features of the carnivorous syndrome such as digestion.

Dansk resumé

Fremskridt i Next Generation sequencing teknologier har gjort det muligt at sekvensere DNA ekstraheret fra miljøer og organismer til en rimelig pris, der har tilladt forskningsfelter som metagenomics og hele transkriptom sekventering (RNA-seq) at blive etableret. Disse teknikker gør det muligt at studere funktionelle relationer i miljøer og også enkelte organismer. Også hidtil ukendte forbindelser og enzymer kan blive udvundet fra sekventeringsdata. Processen med at udnytte de biologiske ressourcer til kommerciel brug er kendt som bioprospektering.

Denne ph.d.-afhandling beskriver konceptet bioprospektering i den post genomiske æra (kapitel 1), og introducerer forskningsfelterne metagenomics og RNA-seq (kapitel 2) som teknikker til at få adgang til og analysere biologiske ressourcer. I bestræbelserne på at opdage og kommercialisere sådanne biologiske ressourcer, skal juridiske forpligtelser være opfyldt, som generelt er reguleret af "Convention on Biological Diversity" (beskrevet i kapitel 3). Proteolytiske enzymer – som er beskrevet i kapitel 4 – er interessant i forbindelse med bioprospektering på grund af deres høje markedsværdi. Del II beskriver metoder, der anvendes til analyse af metagenomiske og RNA-seq datasæt, inkluderet herunder er manuskript I, som beskriver den taksonomiske annotering af et sen pleistocæn heste metagenom og den funktionelle annotation af æsel genomet. Forskning fokuseret på den funktionelle analyse og identifikation af nye proteolytiske enzymer i det polare havmiljø og transkriptom analyser af den kødædende plante *Dionaea muscipula* er også beskrevet.

Polarhavet er et unikt, ekstremt habitat med konstante lave temperaturer og ingen lys indtrængen i dybden. Vandprøver blev indsamlet ved varierende dybde (40 m – 4.300 m) under polarekspeditionerne Galathea III og LOMROG II og blev derefter sekvenseret. Funktionel analyse af polare marine metagenomer kan afdække bakterielle strategier for overlevelse i polarhavet (manuskript II). Desuden er dette ekstreme miljø en rig ressource til udvinding af nye proteolytiske enzymer. Manuskript III præsenterer en bioinformatik tilgang til at identificere sekvenser for potentiel kommercialisering.

Kun få arter i planteriget er i stand til at fange og fordøje byttedyr. Sekventering, assembly og funktionel annotering af et normaliseret transkriptom fra den mest berømte kødædende plante Venus-Fluefanger (*Dionaea muscipula*) er præsenteret i manuskript IV.

Kapitel 12 opsummerer de berørte emner og indeholder afsluttende bemærkninger om de fremtidige perspektiver for den præsenterede forskning. Sammenfattende viser denne afhandling, hvordan biologiske ressourcer kan udnyttes til kommerciel brug. Desuden giver resultaterne en bedre forståelse

af det mikrobielle samfunds overlevelse i det dybhavet. Endelig giver data fra Venus-Fluefanger transkriptomet en tilgængelig ressource til afdækning af egenskaber hos kødædende planter.

Acknowledgements

The work on this thesis would not have been possible without the encouragement of my supervisor Professor Thomas Sicheritz-Pontén, who has always provided advise and guidance when needed. Thanks for sharing your passion and enthusiasm for science with me and giving me the opportunity to be part of exciting research projects.

Without my collaborators, I would not have been able to achieve the goals set for my PhD. I would like to express my gratitude to Søren Sørensen and Lea Skov Hansen from KU Microbiology, John Mundy and Michael Krogh Hansen from KU functional genomics, the research team at GeoGenetics and Jesper Salomon from Novozymes. It has been enjoyable the time working with you.

I was lucky to have been a part of such a great research team, thanks to all members of the Metagenomics group. It has been very inspiring to participate in the scientific discussions at our weekly group meetings, I would especially like to thank Nikolaj Blom, Thomas Nordahl Petersen, Henrik Bjørn Nielsen, Bent Peteresen and Simon Rasmussen for sharing your knowledge with me. Special thanks to Henrik Marcus Geertz-Hansen for making the collaboration with Novozymes easier on my part.

CBS has not just been a place to work, but also to socialize and make friends. Thanks to all of you at CBS for making this working place so special. Thanks to Agata, Arcadio, Kasper, Grace, Dhany, Agnieszka, Juliet and Tejal for creating such a great atmosphere in our office.

Handling big datasets has its challenges, owing to the system administration team's support most technical issues were solved quickly. I would like to thank John Damm Sørensen, Peter Wad Sackett and Kristoffer Rapacki. The CBS administration has always helped me out with formalities, thanks to Lone Boesen, Dorthe Kjærsgaard and Marlene Beck.

I would also like to thank Agata, Asli, Bent, Ida and Tejal for reading through parts of my thesis, giving me input and encouragement during the final stages of the writing process.

Thanks to all of my friends for the moral support. My deepest gratitude goes to my parents, grandmother and siblings for being so very supportive and patient with me during my years of studies. Lastly, to Lise, who has always been there for me and believing in me, thank you so very much.

Papers included in the thesis

- Ludovic Orlando, Aurelien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, Enrico Cappellini, Bent Petersen, Ida Moltke, Philip L. F. Johnson, Matteo Fumagalli, Julia T. Vilstrup, Maanasa Raghavan, Thorfinn Korneliussen, Anna-Sapfo Malaspinas, **Josef K. Vogt**, Damian Szklarczyk, Christian D. Kelstrup, Jakob Vinther, Andrei Dolocan, Jesper Stenderup, Amhed M. V. Velazquez, James Cahill, Morten Rasmussen, Xiaoli Wang, Jiumeng Min, Grant D. Zazula, Andaine Seguin-Orlando, Cecilie Mortensen, Kim Magnussen, John F. Thompson, Jacobo Weinstock, Kristian Gregersen, Knut H. Røed, Vera Eisenmann, Carl J. Rubin, Donald C. Miller, Douglas F. Antczak, Mads F. Bertelsen, Søren Brunak, Khaled A. S. Al-Rasheid, Oliver Ryder, Leif Andersson, John Mundy, Anders Krogh, M. Thomas P. Gilbert, Kurt Kjær, Thomas Sicheritz-Pontén, Lars Juhl Jensen, Jesper V. Olsen, Michael Hofreiter, Rasmus Nielsen, Beth Shapiro, Jun Wang and Eske Willerslev. *Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse*. Nature, 499(7456), pp.74 8, 2013.
- **Josef Korbinian Vogt***, Lea Benedicte Skov Hansen*, Dhany Saptura, Peter Nikolai Holmsgaard, Lars Hestbjerg Hansen, Søren Sørensen, Thomas Sicheritz-Pontén and Nikolaj Blom. *Comparative functional analysis of arctic marine metagenomes reveals strategies for deep sea persistence*. Manuscript in preparation.
- **Josef Korbinian Vogt***, Henrik Marcus Geertz-Hansen*, Lea Benedicte Skov Hansen, Søren Sørensen, Jesper Salomon, Thomas Sicheritz-Pontén and Nikolaj Blom. *Exploiting the polar marine environment for bioprospecting: novel protease discovery*. Manuscript.
- Michael Krogh Jensen*, **Josef Korbinian Vogt***, Simon Bressendorff, Andaine Seguin-Orlando, Hamed El-Serehy, Morten Petersen, Khaled Al-Rasheid, Thomas Sicheritz-Pontén, John Mundy. *Transcriptome and genome analyses of the Venus flytrap (*D. muscipula*)*. Manuscript ready for submission.

* These authors contributed equally.

Papers not included in the thesis

- Nitsch, D., Tranchevent, L.-C., Goncalves, J.P., **Vogt, J.K.**, Madeira, S.C. and Moreau, Y. *PINTA: A web server for network-based gene prioritization from expression data*. Nucleic Acids Research, 39 (suppl 2): W334-W338, 2011.

Part I

Introduction

Chapter 1

General Introduction

This PhD thesis focuses on the functional analysis of genomes and environmental samples and how biological resources can be exploited for industrial use.

Functional analysis of environments and single organisms reveals strategies how organisms adapted to its habitat and why specific traits have been developed.

Research fields applying Next Generation sequencing technologies, such as metagenomics and whole transcriptome sequencing (RNA-seq) analysis, aid in accessing the encoded functions of the DNA or its transcripts.

The research field metagenomics makes it possible to directly sequence and analyze environmental samples without the need for cultivation. In connection with the functional analysis efforts of metagenomes, the donkey genome was functionally annotated and the taxonomic composition of a Middle Pleistocene horse was investigated. The results were a part of manuscript I which will be presented in the methods part as a lot of principle knowledge about NGS data handling and taxonomic annotation was acquired. RNA-seq analysis aims to research transcribed genes within a genome. An elaborate explanation of metagenomics and RNA-seq is provided in Section 2.1 and Section 2.2 respectively.

The sequencing data can also be used to explore and identify novel compounds for commercialization. Tapping into natural resources is also known as bioprospecting. Bioprospecting is an application-driven strategy for utilizing natural resources for industrial use. The principle workflow of a typical bioprospecting project is shown in Figure 1.1. As a first steps, probes are sampled from an (extreme) environment and sequenced. The sequencing data are analyzed and potential targets for commercialization are identified. Sampling and exploitation of such resources is governed by the Convention on Biological Diversity (CBD) [33]. The ethics behind bioprospecting and

governance of environmental samples is explained in Chapter 3.

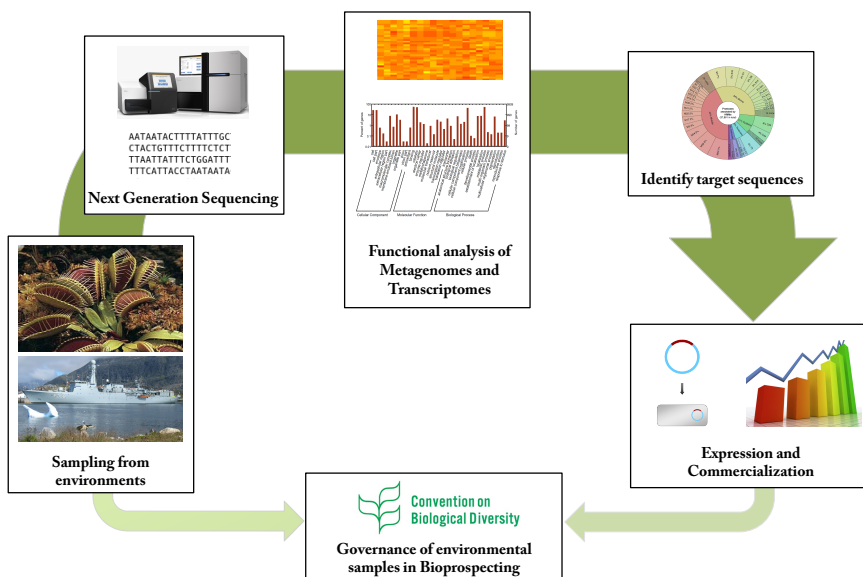


Figure 1.1. Illustration of bioprospecting and functional analysis of environmental samples. As a first step environmental samples are gathered and sequenced. The sequencing data is analyzed and target sequences are identified. Functional analyses of the samples gives insight in the environment or single genome and can aid the identification of possible targets. Bioprospecting also includes expression of the promising targets together with possible commercialization. Analyzing and commercializing of most environmental samples are governed by the convention on biological diversity, i.e. regulation of commercial use of metagenomic samples. Illustrations were adapted from Illumina (<http://www.illumina.com/>, accessed 1. November 2013.) and the Convention of Biological Diversity homepage (<http://www.cbd.int/>, accessed 1. November 2013.).

The research projects presented in this thesis are based on discovering environments, which exhibit protease activity such as (1) microbiological environments where proteolytic enzymes are adapted to environmental conditions of interest to the biotech industry (high pH, high pressure and low temperature), and (2) unexplored environments, rich in proteolytic enzymes. Proteases are enzymes which are capable of cleaving other proteins (amino acid chains) or even themselves in catalytic fashion. This enzyme class exhibits a high market value. Chapter 4 elaborates on the properties and applications of proteases.

As mentioned earlier, bioprospecting is generally governed by the CBD. However, research with the aim of commercialization can be hindered due to royalty issues and compensation payments. Metagenomic samples of the Ikka columns in Greenland were initially part of the project due to the environment's constant low temperature and high pH, they would have been an optimal target for finding novel proteases [20, 143]. Even though, exploitation of biological resources from Greenland is regulated by the CBD and the law on Commercial Exploitation of Greenlandic Biological Resources, no agreement could be reached between the involved parties for a reasonable commercialization of the natural resources. Thus, the Ikka column samples were replaced with water samples of the polar marine environment collected during the Galathea III and LOMROG II polar expeditions as these samples comply with the CBD and do not subject to national law for financial compensation. The samples span the entire water column from 40 m – 4,300 m. Bacteria such as *Pseudoaltermonas* [184] from deep-sea environments have already been shown to produce pressure-stable proteolytic enzymes. In addition to pressure-stability, several of the enzymes from the deep-sea are also known to be active at cold temperatures [184].

Furthermore, the transcriptome of the carnivorous plant Venus flytrap (*Dionaea muscipula*) was analyzed. The Venus flytrap is a promising target for identifying proteases as the digestive fluid of carnivores plants is capable of digesting an intact animal (e.g. a fly) as the only representative of the plant kingdom. The proteolytic enzymes in these plants may exhibit properties and activities that differ from other organisms. The transcriptome data will aid in identifying proteases in future analysis.

Chapter 2

Next Generation Sequencing

Since the advent of the Human Genome Project, new sequencing technologies arose, which expand the applications of sequencing data. New analysis methods make it possible to utilize the genomic sequence information in various research projects and in the industry. High throughput or Next Generation Sequencing (NGS) makes sequencing of genomes reasonable at low costs and high coverage. The sequencing costs dropped dramatically within recent years (Figure 2.1), making it feasible to include sequencing in numerous biological experiments [173]. NGS opens the door for many types of analyses such as metagenomic studies, sequencing of the transcriptome and many more [81, 110, 151]. The following sections will introduce the concept of metagenomics and whole transcriptome shotgun sequencing. Several sequencing technologies are available with different specific advantages and disadvantages. Table 2.1 gives an overview of the most predominant NGS sequencing technologies together with their specification, advantages and disadvantages.

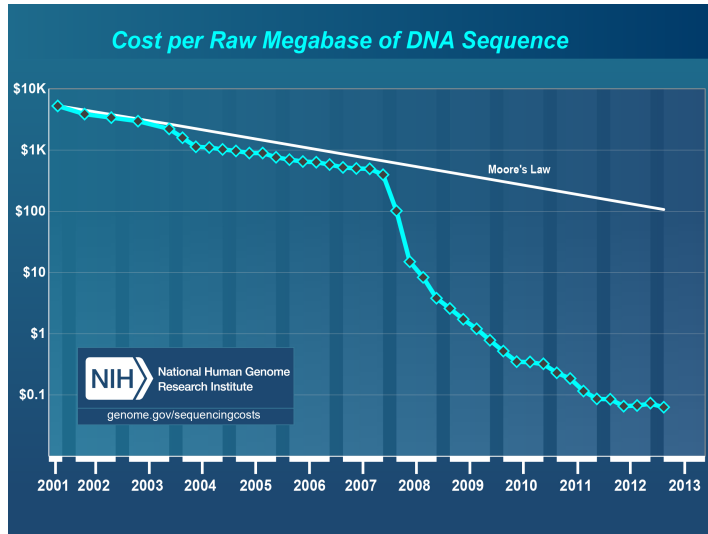


Figure 2.1. Sequencing cost per raw megabase of DNA. The development in sequencing costs started to outperform Moore's law the years 2007 – 2008. Illustrations adapted from the National Human Genome Research Institute (www.genome.gov/sequencingcosts, accessed 15. October 2013.).

Table 2.1. Overview of next-generation sequencing methods [104, 110, 151, 103, 132, 51]. Specifications were adapted from Pacific Biosciences, life technologies (<http://www.pacificbiosciences.com/>), <http://www.appliedbiosystems.com/>, <http://www.illumina.com/>, accessed 14. October 2013.

Method	Pacific Bio	Ion Torrent	Pyrosequencing, Illumina 454	SOLID	Sanger sequencing
Read length	> 5,000	< 400 bp	< 700 bp	50+50 or 50+35 bp	< 900 bp
Reads per run	50 thousand/- cell	< 80 million	1 million	< 1.4 billion	N/A
Time per run	< 2 hours	2 hours	24 hours	27 hours - 11 days	< 3 hours
Cost per 1 million bases (in US\$)	< \$1.50	\$1	\$10	\$0.05 to \$0.15	\$2,400
Accuracy	87% (read length mode), > 99% (accuracy mode)	98%	99.9%	98%	99.9%
Advantages	very long read length, fast	inexpensive equipment, fast error rate	long reads, fast	high sequence yield	long individual reads
Disadvantages	low yield at high accuracy; expensive equipment		expensive runs	expensive equipment	expensive, not suitable for larger sequencing projects

2.1 Metagenomics

It is estimated that the earth hosts $>10^{30}$ microbial cells [179]. This figure exceeds the number of known stars in the universe by nine orders of magnitude. This richness of single-celled life, the first life to evolve on the planet, still accounts for the vast majority of functional drivers of our planet's ecosystems [47] but the diversity and interdependencies of these microscopic organisms remain largely unknown [88].

Cultivation is the most important laboratory technique in conservative microbiology. The majority of microbial life, however, cannot be discovered with these traditional laboratory (cultivation) based approaches. Amann et al. [5] estimated that only 1% of environmental microbial species can be readily cultivated, whereat the remaining 99% are not accessible for research. Metagenomics, however, compasses the need for cultivation. Metagenomics can be described as both, (1) a set of research techniques and (2) a research field [62]. Its main goal is to study DNA and translated protein sequences from all the genomes found in an environment [168]. Metagenomic samples originate from "common" or extreme environmental habitats, such as sea water from the Sargasso Sea [172], acid mine drainage [133], whale fall [167] or the human gut [118].

Methods for environmental sampling vary depending on the purpose of the study, the habitat sampled and the desired downstream analysis. In principle three different approaches to metagenomics analysis exist. Figure 2.2 depicts a typical metagenomics workflow. The clone library approach (Figure 2.2 left), where random DNA amplification and cloning into host vectors is followed by host expression, focuses on expression of the amplified DNA in a suitable host. Subsequent screening for e.g. enzymatic activity – functional screening – is the major goal in this kind of analysis. Companies, such as Verenum apply this method to systematically screen metagenomes for novel catalytic proteins or improved products¹.

The amplicon library approach (Figure 2.2 middle) targets specific genes, such as the 16S rRNA gene. Those genes are enriched by PCR. The library is sequenced and the typical abundance (OTU abundance) is analyzed.

The third high throughput approach (Figure 2.2 right) does usually not include an intermediate gene selection and amplification step before sequencing. However, depending on the sequencing technology, the sequencing process itself might include an amplification step. The isolated metagenomic DNA is then sequenced directly and usually to a very high depth using a high throughput platform. The output sequences which are referred to as "reads" are then assembled. Thereby, a library of contigs is created which serves as starting point for computational analysis, such as gene annotation, sequence translation and functional annotation.

Despite the fact that the science of metagenomics as an accredited research field is only a few years old, a lot of interest into the emerging scientific field

¹<http://www.verenum.com/>, accessed 1. October 2013.

has been observed. For the period of 2005 to 2008, 18 review articles and 45 original research papers were published [85]. A SciVerse Scopus search on metagenomics in the period of the beginning of 2005 to September 2013 already results in 2,739 review and original research papers, where approximately 50% of all metagenomics papers have been published in the years 2012 and 2013. Still, new computational methods and pipelines are evolving in order to tackle such huge amounts of data. These techniques should serve the need of maximizing the understanding of the genetic diversity and activities within the sampled community. As technologies for DNA extraction, sequencing, assembly and annotations are constantly optimized, the potential to generate complete genome data, reconstruct large DNA fragments and assign functions to genes and translated amino acid sequences have greatly improved.

A high throughput sequencing approach poses a challenge for downstream bioinformatics analysis. Extracting meaningful information from the millions or billions of genomic sequences is challenging for bioinformaticians. Where assembly and annotation of sequencing data from a single cultured organism is a manageable task, handling of metagenomics data is far less trivial. In metagenomics, the data come heterogeneous microbial communities, at times comprised of more than 10,000 species, with the sequence data being noisy and mostly partial [182]. It is difficult to capture the whole truth or composition of a community.

Within the last decade, the number of public metagenomes is ever increasing. Metagenomes can be retrieved from online databases collecting various metagenomic data sets. Such databases are the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) [150] or Genomes Online [14]. The latter is an online database for comprehensive access to information regarding complete and ongoing genome projects, as well as metagenomes and metadata, around the world. As by May 2013, the Genomes Online Database reported over 376 metagenome studies in progress with 2,749 samples being researched¹. CAMERA makes raw environmental sequence data, partial assemblies, genes, associated metadata, precomputed search results, and high-performance computational resources accessible for a broad research community.

¹<http://www.genomesonline.org/cgi-bin/GOLD/index.cgi>, accessed 20. May 2013.

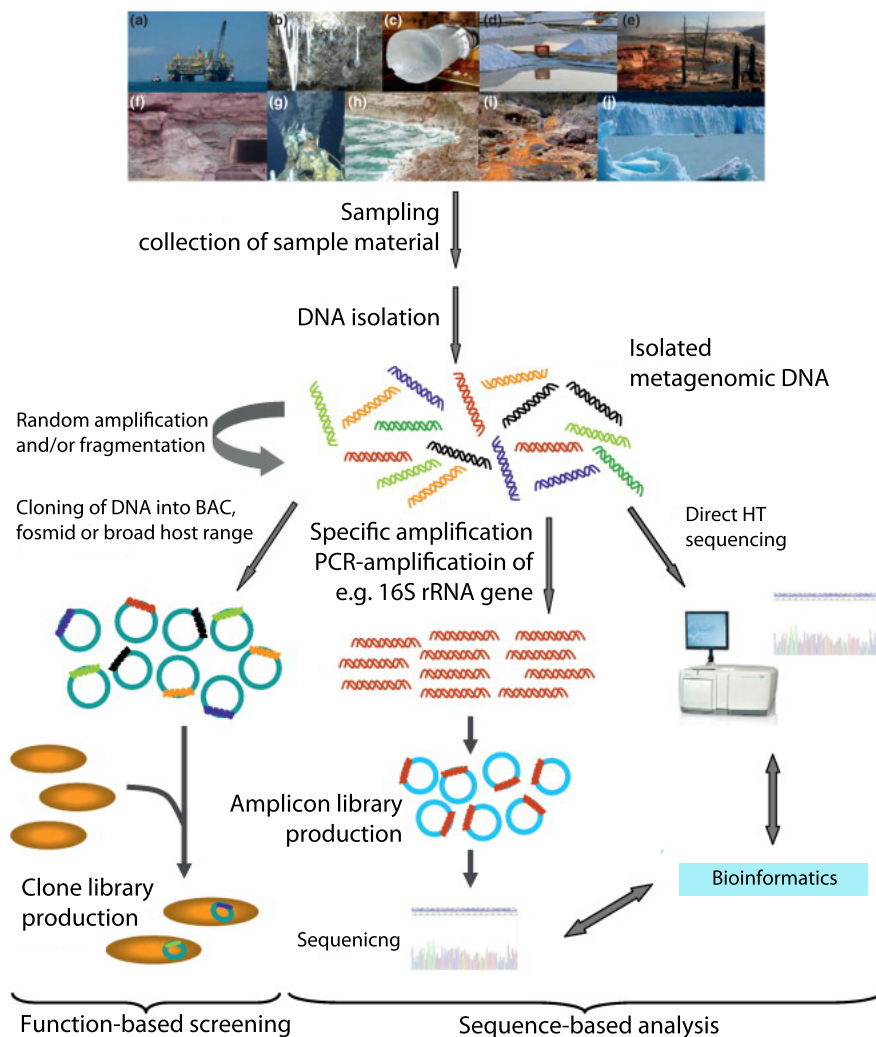


Figure 2.2. A typical metagenomic experiment starts with sampling and DNA isolation. The DNA is fragmented and can either be directly sequenced and analyzed or the fragments can be cloned into vectors and screen. 16S analysis of metagenomic samples include a PCR amplification of 16S rRNA genes; (a) offshore oil reservoir, (b) acidic snotties, (c) ice core, (d) saltern crystallizer ponds, (e) Yellowstone national park hot spring, (f) mine, (g) black smoker chimney, (h) the Dead Sea, (i) and (j) glacier ice. Illustration modified from Lewin [96].

2.2 Whole Transcriptome Shotgun Sequencing

The transcriptome represents the total set of RNA in a given organism [171]. In contrast to a genome, which can be described as fixed within a cell, the transcriptome varies within time. The transcriptome mostly gets described as the genes that are being actively expressed by the cell at the point of analysis, i.e. mRNA within the cell. Technically, however, mRNAs only cover a subset of a transcriptome [171].

Earlier methods for transcriptome analyses such as microarray and qPCR based technologies require prior knowledge. *A priori* knowledge, however, does not exist for many organisms. Whole Transcriptome Shotgun Sequencing (RNA-seq) does not require knowledge *a priori* such as well-characterized species. It takes advantages of the advances within next generation sequencing to sequence the entity of RNAs present within the cell making high-throughput sequencing technology the standard method for accessing the transcriptome of an organism [134]. Its applications range from expression profiling to differential gene expression studies [81, 116, 121].

Figure 2.3 depicts a typical library preparation workflow for an RNA-seq experiment. RNA library preparation starts with RNA (mRNA) extraction, followed by removal of DNA contamination using DNase. The remaining RNA is fragmented into short segments. In contrast to genomics or metagenomics sequencing approaches, RNA-seq setups involve a cDNA synthesis step, where the RNA fragments are reverse transcribed into cDNA. The cDNA is sequenced after adaptor ligation and fragment size selection.

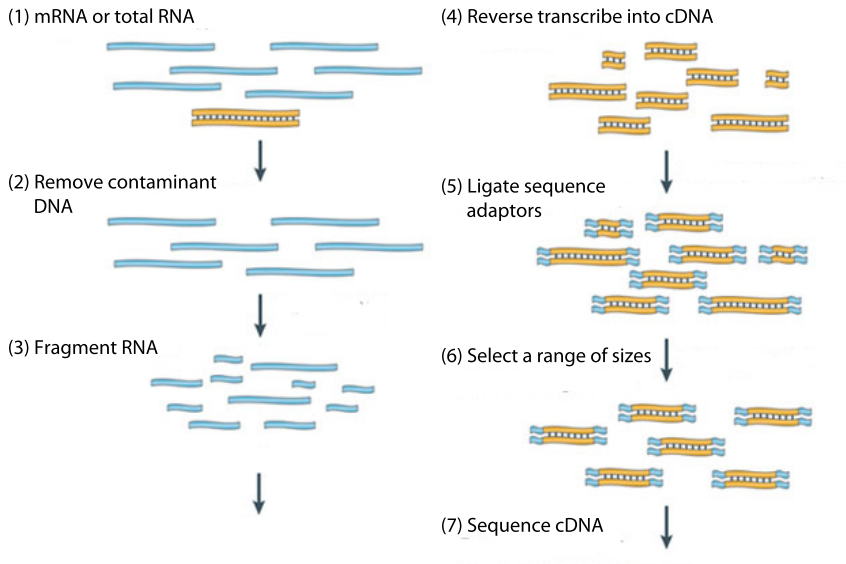


Figure 2.3. Workflow of an RNA-seq experiment. (1) Total or mRNA is extracted, (2) DNA contamination is removed, (3) fragmentation of RNA, (4) reverse transcription, (5) sequence adaptor ligation and range selection (6), final sequencing of the cDNA. Illustration adapted from Martin and Wang [107].

Chapter 3

Governance of Environmental Samples

This chapter provides an introduction to the ethics behind the use of metagenomic data sampled from environments in academia and in the industry. This section, however, neither discusses biopiracy in context of directed exploitation of indigenous knowledge, nor the implications arising from patents in an agricultural context.

Metagenomics as a research field strives to unravel the diversity of ecosystems, describing dependencies of species within the system and hunting for novel compounds, enzymes and pathways. Such findings can bare a huge potential for industrial applications. However, one has to be certain about the origin of the data. Sample sources or the intellectual properties of the samples have to be secured. An extensive debate about the ethics and intellectual property rights of natural sources being exploited by academia and industry started in the late 90s and the early 2000s. The academic dispute pinpointed biopiracy and the exploitation of natural material for bioprospecting as the major cause of the conflict. In order to grasp the implications of biopiracy, one has to understand the concept of bioprospecting.

3.1 Bioprospecting as a research strategy

Metagenomics and bioprospecting share many characteristics given their aim to uncover novel characteristics from environments. While metagenomics is a defined research field, bioprospecting is coined as a process encompassing several techniques but more importantly it is an application-driven strategy. However, both are very much interlinked and can also be described as the opposite sides of the same coin.

Bioprospecting originates from the field of chemical ecology dating back to the late 1950s [64]. Chemical ecology is the study of chemical compounds directly associated with the interactions between organisms and their occupied environment [64]. The research of natural products increased considerably followed by a directed "pursuit" of commercializable natural compounds - eventually establishing chemical prospecting [43]. While similar in principle, chemical prospecting solely relied upon chemical synthesis of newly discovered, commercially-relevant compounds, bioprospecting in the post genome area takes advantage of decreased sequencing cost, metagenomics as screening technology and advanced bioinformatics infrastructures.

An up-to-date definition of bioprospecting is divergent. It covers stages of searching and sampling of resources to be used in applications and development, the common understanding of bioprospecting, however, is the commercial exploitation of such research. It includes the following elements¹:

- systematic search, collection, gathering or sampling of biological resources for purposes of commercial or industrial exploitation
- screening, isolation, characterization of commercially useful compounds
- testing and trials
- further application and development of the isolated compounds for commercial purposes, including large-scale collection, development of mass culture techniques, and conduct of trials for approval for commercial sale

3.2 Biopiracy

Since bioprospecting became an important research concept, the matter of property rights, ethics and exploitation became an issue pointing out biopiracy as the main reason. Biopiracy has various definitions, the most coherent definitions of biopiracy in the context of metagenomics are as follows:

(1) "Biopiracy can be defined as the intentional theft of indigenous and traditional knowledge and resources of indigenous people for commercialization (and profit) without permission, recognition or compensation to the indigenous peoples from which it originated" [111].

(2) "Oftentimes biopiracy is described as the commercial development of naturally occurring biological materials [...] by a technologically advanced country or organization without fair

¹adapted from the United Nations University's report: Bioprospecting of Genetic Resources in the Deep Seabed: Scientific, Legal and Policy Aspects, <http://www.ias.unu.edu/binaries2/DeepSeabed.pdf>, accessed 20 September 2013

compensation to the peoples or nations in whose territory the materials were originally discovered”¹.

The Merck-INbio Agreement² is one of the most prominent examples of a case where biopiracy efforts have been everted by compensating the country where the natural resource originates [32].

In order to address biopiracy the Convention on Biological Diversity (CBD) was established, providing guidelines for biological resource administration, i. e. metagenomic datasets for bioprospecting [33, 70].

3.3 Convention on Biological Diversity

The Convention on Biological Diversity (CBD) was signed by 150 governments, including the United States, at the 1992 Rio Earth Summit. The agreement was later ratified by more than 187 countries, however not the United States, and is an international legally binding treaty. The treaty is in charge of administrating and discussing issues related to the CBD³.

The CBD ”[...] recognizes that biological diversity is about more than plants, animals and microorganisms and their ecosystems, it is about people and our need for food security, medicines, fresh air and water, shelter, and a clean and healthy environment in which to live”⁴. In summary the convention has three major goals [62]:

- conservation of biological diversity (or biodiversity)
- sustainable use of biological components
- fair and equitable sharing of benefits arising from genetic resources

States gain sovereignty over their own natural resources by subjecting its distribution to their national legislation. The treaty also grants protection of (1) intellectual property rights to its nation and (2) compensation or benefit-sharing from commercialized findings.

In the context of metagenomics, it is important to recognize that metagenomics projects rely foremost on collecting environmental samples. The collection within national borders is strictly guided by the CBD. Sample collection outside national borders, however, are not part of the CBD, e.g.

¹quoted from the Free Dictionary, <http://www.thefreedictionary.com/biopiracy>, accessed 25 September 2013

²Against payment of a certain sum of money and profit-sharing, the U.S. American company Merck has been granted a temporary right to perform pharmacological analysis of genetic resources in Costa Rica, as well as the right to patent the developed drugs. In exchange, it had to provide the INBio laboratory with scientific equipment. A major part of the money was invested in the conservation of Costa Rican national parks, so that, in this case, the utilization of biological diversity eventually contributes to its conservation [32]

³<http://www.cbd.int/convention/>, accessed 26 September 2013

⁴quoted from <http://www.cbd.int/convention/>, accessed 26 September 2013

deep-sea vents beyond national jurisdictions or the [62]. Territories which are not under national jurisdiction may, however, be regulated under different agreements, e.g. Antarctica. Antarctica is regulated under the Antarctic Treaty System (ATS) from 1961, signed by 50 nations. The treaty allows organisms to be taken, patented and commercialized [67]. Large scale extraction of organisms is not permitted to limit adverse impacts on the Antarctic environment and dependent and associated ecosystems¹.

Compliance to the CBD nowadays is very important to prevent charges of biopiracy. Most companies which are specialized in commercializing metagenomic driven research results declare their compliance to the CBD as advantages, such as Novozymes² or Verenium³. The latter markets itself as "pioneer in the field of ethical bioprospecting", profiting from compliance to the CBD by integration into the corporate identity.

¹<http://www.ats.aq/>, accessed, 30 September 2013

²retrieved from Novozymes' annual report 2010<http://report2010.novozymes.com>, accessed 19. September 2013

³<http://www.verenium.com/ourwork3.html>, accessed 19. September 2013

Chapter 4

Proteolytic enzymes

The bioprospecting efforts in this research focused on proteolytic enzymes, the following chapter introduces this enzyme class. Proteolytic enzymes (also termed proteinases, peptidases or proteases) are enzymes which are capable of cleaving other proteins (amino acid chains) or even themselves in catalytic fashion. They make-up the largest single family of enzymes [100]. Through structural and functional diversity, proteases carry out a vast array of critical functions ranging from intracellular protein recycling to nutrient digestion, immune system cascade amplification, signal transduction and also blood coagulation [66].

4.1 Classification

Proteases are divided into broad groups according to their catalytic abilities. Each peptidase can be assigned to an Enzyme Commission number (EC number) [177]. The EC number is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze [177]. Proteases are classified into aminopeptidases, dipeptidases, dipeptidylpeptidases, peptidylpeptidases, carboxypeptidases and endopeptidases according to the reaction site [10]. Figure 4.1 gives an overview of the enzymatic reactions.

There are, however, several limitations to the EC number classification, since it does not reflect evolutionary relationships. Related peptidases can have identical substrate specificities [128]. The mechanism used to cleave a peptide bond involves making an amino acid residue that has the cysteine and threonine (proteases) or a water molecule (aspartic acid, metallo- and glutamic acid proteases) nucleophilic so that it can attack the peptide carboxyl group [128]. One way to make a nucleophile is by a catalytic triad, where a histidine residue is used to activate serine, cysteine, or threonine

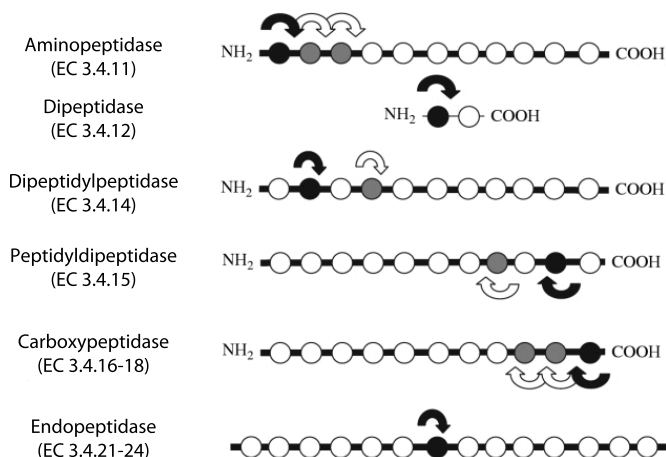


Figure 4.1. Classification of peptidases by the catalyzed reaction. Beads represent amino acids, string the peptide bonds. Black arrows indicate the first cleavage and white arrows subsequent cleavages. For the first cleavage, the amino acid(s) to which specificity is mainly directed is shown in black and for subsequent cleavages in grey. Illustration adapted from Polaina and MacCabe [128].

as a nucleophile [128]. Threonine and glutamic-acid proteases were first assigned to its own family by 1995 and 2004, respectively [135, 136]. Rawlings and Barrett [135, 136] classified proteases within those broad groups into families of associated proteases. For instance the serine protease family proteases are assigned to a Sx label - where S describes the serine catalysis and x the corresponding family association, e.g. S1 stands for chymotrypsin. Table 4.1 summarizes the families and its most known representatives.

The most abundant resource for protease family classification is found in the MEROPS database¹ that has been growing considerably within the last years. A 10-fold increase of deposited sequences has been observed within the last decade [136]. The database provides a hierarchical classifications in which homologous sets of peptidases and protein inhibitors are grouped into protein species, which are in turn clustered into families and into clans. Families are assigned on the basis of statistically significant similarities in amino acid sequence, and families which are homologous are clustered into clans [135, 136].

¹<http://merops.sanger.ac.uk>, accessed 1. October 2013.

Table 4.1. Protease families, highlighted letters denote the Merops family [135, 136]. Amino acid (AA), active center (AC)

Protease family	Funct. AA or AC	Example
Aspartic	Aspartic acid	Pepsin, Chymopepsin
Cysteine	Cysteine	Papain, Cathepsin K, Caspase
Glutamic	Glutamic acid	scytalidoglutamic peptidase
Metallo	coordination complex	Collagenase, Carboxypeptidase A and B
AsparagiNe	Asparagine	virus Lyases and Coat proteins
Serine	Serine	Trypsin, Plasmin, Thrombin
Threonine	Threonine	various Acetyl and glutamyl-transferases
(Mixed/P)	-	DmpA aminopeptidase
Unknow	-	Collagenase

4.2 Proteolytic enzymes in the industry

Proteases, their substrates and inhibitors are of great relevance to biology, medicine and biotechnology and are oftentimes used in industrial setups. They are particularly sought after their hydrolyzing peptide bond capabilities in aqueous environments and also the peptide bond synthesis in non-aqueous biocatalysis. A selection of applications is shown in Table 4.2.

Table 4.2. Industrial applications of proteases [191, 87]

Industry	Applications
Detergent	Protein stain removal (laundry and dish wash)
Starch and fuel	Protease (yeast nutrition fuel)
Food	Flavor, milk clotting, infant formulas (low allergenic)
Pulp and paper	Biofilm removal
Leather	Unhearing, bating

Proteolytic enzymes are already estimated to account for more than 60% of the total sales on the enzyme market worldwide with an annual sales worth of about 1.5 – 1.8 billion US dollars [76, 176]. Detergent proteases alone, with an annual market of about 1 billion US dollars account for the largest protease application segment [176]. The demand for proteolytic enzymes is still rising; the compound annual growth rate (CAGR) is estimated at 7.2% from

2013 to 2018¹.

Thus, proteases are a profitable class of enzymes to be studied, especially sequences originating from extreme environments. Next generation sequencing technologies are useful tools to access these environments. The concept of next generation sequencing and its applications are described in Chapter 2.

¹<http://www.prweb.com/releases/protease-leads-feed/enzyme-market-phytase-nsp/prweb10754350.htm>, accessed 10. September 2013

Part II

Methods

Chapter 5

From Sequencing Reads to Sequence Assembly

Sequence assembly is an essential step when analyzing sequencing data from for example metagenomes or transcriptomic studies. The primary goal of a sequence assembly lies in the construction of contigs which are longer than the sequencing reads provided from a sequencer. Prior to an assembly process or any kind of sequence data analysis, it is common practice to pre-process the provided sequencing data.

5.1 Pre-Processing of Sequencing reads

Raw sequencing data is usually delivered in FASTQ format as shown in Figure 5.1.

A FASTQ sequence block consists of four distinct lines starting with a header denoted by @ followed by the raw read sequence in the next line. The third line, which is indicated with +, is an optional Illumina header. FASTQ

```
@HWI-ST575:107:C0HE6ACXX:4:2307:10548
AGAAGATGCCCTGGGTGCCGTTGCCATGGTGGGAAGTC
+
@@@FD>DDBFBBHGG<<<A8<C;?F>?F;1:CGGCDD
@HWI-ST575:107:C0HE6ACXX:4:2307:10622
GGTCATGGCGTCTTACTGCCTTTCGCCACCGATCCGT
+
18=?+B+: :)0@F+C2C,+A+2<+A)@EDH:?6?FG6
```

Figure 5.1. Example of a FASTQ formatted file

formatted sequencing reads are accompanied by ASCII character probability scores in the fourth line. The probability score is the so called Phred quality score and is a non-negative integer that describes the error probability of a base call to be wrong [30, 45, 46]. The error score Q can be written as the following:

$$Q_{phred} = -10\log(err)/\log(10) \quad (5.1)$$

err : probability of a wrong base call

For instance, to allow for an error probability of <0.01 , a minimum Q score of 20 is required. Even though a Q score of 20 is mostly used, a 1% error in millions of base pairs still adds up to a considerable amount of errors. For most applications an error probability of <0.01 is sufficient. To assure correct base calling, the Phred scores are used to determine an accumulated cutoff where the trailing part of the read sequence is trimmed as the base qualities decrease towards the end of the read.

The maximum sequencing read length of for example the Illumina HiSeq 2000 sequencer are commonly 2 x 100 base pairs when running in paired-end mode¹. A detailed overview of sequencing technologies and generated read lengths is provided in Table 2.1. However, erroneous nucleotides in the sequence have to be accounted for. The tool FastQC [144] is normally used for quality assessment.

We used a custom python script to trim raw reads of the polar marine metagenomes and the RNA-seq data from the Venus flytrap, so that the minimum Phred score cutoff for individual bases and average read score were set to 20. Reads with a minimum length of 35 nucleotides were used for further processing. To assure consistency within sequencing datasets, it is also good practice to investigate for deviating GC content, overrepresented sequences and k -mers. However, k -mer correction (k -mers are sub-sequences with a fixed length k ; the concept of k -mers is explained in more detail in Section 5.2) in metagenomic datasets is not as trivial as in single genome datasets (commonly done with Quake [84]) as it discards low coverage data detected as low-abundance k -mers. No separate error correction step was included in the analysis of the polar marine metagenomic samples (Manuscript II and III) as the assembly method used can correct for reads in regions with high depth. Furthermore, no error correction was applied on the transcriptome data of the Venus flytrap (Manuscript IV) as at the time of analysis no appropriate tool was available which improved assembly quality. However, designated methods for RNA-seq error correction are being established, such as the HMM-based correction tools SEECER [94], which could improve downstream analysis in the future.

To avoid contamination in sequencing reads it is common practice to remove obvious contamination source DNA, e.g. human DNA. In the presented

¹http://www.illumina.com/systems/hiseq_comparison.ilmn, accessed 5. November 2013

studies, we removed human DNA by mapping trimmed reads to the human genome with BWA [97]. Unmapped reads were used for further analysis.

5.2 *De novo* Assembly

As stated in Chapter 2, next generation sequencing technologies generate a vast amount of short-read sequences. However, making sense out of these short snippets of DNA proves to be challenging. The interest in sequence and genome assembly is ever increasing, especially the need for computationally efficient computing techniques. Furthermore, sequencing applications, such as gene expression analysis, discovery of genomic variants and metagenomic studies, have different requirements for genome assemblies. This section summarizes the de Bruijn graph principle, common assembly tools and their applications.

The de Bruijn Graph Principle

De novo assembly encompasses the need of reference genomes and annotations for genome assembly. Most assembly methods are based on the de Bruijn graph method. This approach considers sequencing reads not as an entity but as a string comprised of multiple *k*-mers [26]. The nucleotide string in Figure 5.2 is cut in *3*-mers ($k=3$) where each consecutive *k*-mer frame is moved to the right. The sequence is then represented as *k*-mers (Figure 5.2 A) with nodes and branches. The path through the graph is then condensed around the node sequences (colored in red). Therefore, redundancies within the overlaps are reduced and computation of the path is more feasible [31, 106]. Sequencing reads are assembled into contigs. Due to repeat regions or erroneous reads, many branches can be introduced in the de Bruijn graph especially when k is small (so called branching problem) [124]. Thus, choosing a proper k value is a crucial tradeoff as short *k*-mers lead to fewer gaps but more branches, while longer *k*-mers lead to fewer branches but more gaps [124].

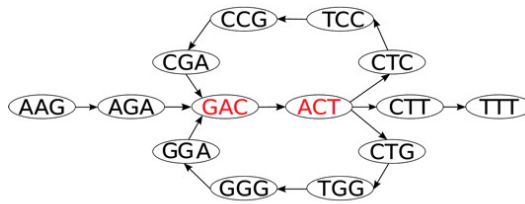
Many assembly tools also take paired-end information into account during the assembly process [59, 124, 153]. Paired-end information makes it possible to close gaps between contigs. Contigs that can be connected with paired-end information are called scaffolds.

The reduced computational complexity as a result of decreased redundancy makes this approach the method of choice for various *de novo* assembly tools. Table 5.1 gives an overview of popular assembly tools for Illumina sequencing reads together with their scope of application.

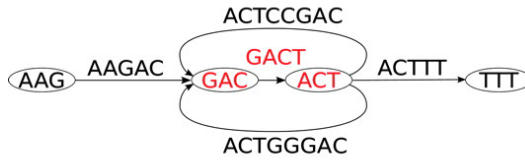
Metagenomic assembly

Next generation sequencing reads of metagenomic samples comprise a vast selection of different species. Reference based assemblies [58], where the reads are mapped to a known genome are not applicable since a high fraction of species in the samples are unknown [182]. Further complications with

AAGACTCCGACTGGGACTTT



A: de Bruijn graph of a sequence



B: condensed de Bruijn graph

Figure 5.2. Schematic illustration of the de Bruijn graph principle in genome assembly ($k=3$); (A) the nucleotide sequence is represented as k -mers, (B) condensed graph representation around the nodes (red). Illustration modified from Chaisson et al. [26].

Table 5.1. Popular *de novo* assembly tools; *used in Manuscript II and III; **used in Manuscript IV

Tool	Application	Reference
ABYSS	metagenome/genome	[153]
Allpaths-LG	(large) genome	[59]
Cufflink	transcriptome	[137]
Meta-IDBA*	metagenome/genome	[124]
MetaVelvet	metagenome	[117]
Oases**	transcriptome	[145]
SOAPdenovo2	metagenome/genome	[105]
Trinity	transcriptome	[60]
Trans-ABYSS	transcriptome	[138]
Velvet**	genome	[190]

metagenomic assembly arise from genomic variance in natural populations [141, 180], poor community coverage [182] and also the risk of chimeric sequence creation [27, 109, 127]. Additionally, the use of repetitive reads appear to be counter-intuitive [71]. It is advised to remove repetitive reads from the assembly of clonal genomes as the de Bruijn graph construction is hindered (e.g. Abyss, Velvet). When assembling metagenomes, however, repetitive reads are likely to originate from dominant species and are advised to be assembled together [164].

Hence, to obtain the best possible assembly, it is important to optimize the assembling process before proceeding with the downstream analysis. More accurate and longer contigs would improve the study of metagenomic datasets such as binning, gene prediction or functional annotation [124]. There are designated metagenome assemblers available, such as MetaVelvet [117] and IDBAs metagenome assembler [124]. All of these tools are based on the de Bruijn graph principle as described earlier [126].

For the analysis of the arctic marine environment metagenomes (Manuscript II and III), the gold standard at that time was the Meta-IDBA metagenome assembler. In addition, Meta-IDBA outperformed SOAPdenovo in terms of assembly quality. The 26 metagenomic samples were individually assembled with Meta-IDBA using the paired-end mode. The sample assemblies were run with varying *k-mers*. Instead of using a singular threshold, Meta-IDBA runs multiple depth-relative thresholds to delete suboptimal *k-mers* in regions with low and high depth. This technique of local assembly with paired-end information is used to solve the branching problem of low-depth short repeat regions [124].

Commonly, metagenomic assemblies do not have full coverage of all organisms in the environment, since sequencing rarely produces all the sequences required for a complete assembly. Therefore, it is important to keep in mind that the observed diversity in metagenomic datasets is not a 100% representation of the full environment.

Transcriptome assembly

Sequencing RNAs or so called RNA-seq has made a big impact on the field of transcriptomics [16, 175]. However, reference genomes required in *ab initio* methods, are not available for numerous organisms [38, 77, 114, 160, 166, 187], e.g. the Venus flytrap. Thus, *de novo* assembly methods were used to achieve assembly of these novel organisms (Manuscript IV). In contrast to metagenomic assemblies, RNA-seq assemblies of clonal eukaryotic organisms do not face the problem of multi-species samples. The extracted RNA is amplified as cDNA and sequenced (see Figure 2.3 in Chapter 2) and it can be assumed that after removing contamination from pre-processed sequencing reads, only target cDNA is present.

Detection of splicing variants is the major challenge in transcriptome assemblies as multiple copies of a gene are transcribed with a varying exon pattern [23, 61, 65]. Designated RNA-seq assembly methods are available with options for splicing variant calling. Table 5.1 gives a short overview of popular

assembly tools.

The *de novo* assembly tool Velvet/Oases [145] was chosen for the Venus flytrap analysis as it was the method of choice at the time of analysis. Pre-processed sequencing reads are provided to a multi *k-mer* assembly done by Velvet. Oases handles transcription variants (isoforms of a transcript) by solving branching points in the de Bruijn graph as loci.

5.3 Assembly assessment

To distinguish between poor and well assembled genomes it is important to get a measure of the assembly quality. However, the quality of assemblies can vary greatly from single genome assemblies to metagenomic assemblies and transcriptome assemblies. When assembling single genomes, one aims for a closed genome, i.e. one contig per chromosome or plasmid, to reach the standard of a High-Quality Draft [25]. This is difficult to achieve when assembling metagenomes due to the fragmented data and the numerous organisms in the samples. Assemblies of cDNA would in the best case result in full-length transcripts.

In order to get a sense of the assembly, various measures are calculated. The *N50* measure is the most widely used measure for assessing the assembly quality. The *N50* is a statistical measure of average length of a set of sequences¹. It can be explained as following: contig or scaffold *N50* is a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value [139, 174]. Thus, the greater the *N50* value, the better the assembly. However, the number of contigs or scaffolds in the assembly and length of the longest contig or scaffold is also an important indicator for the assembly quality. The "assemblathon stats" script was used to calculate assembly metrics² after each assembly in projects presented in this thesis.

Assessing transcriptome assemblies

In transcriptome assemblies, sizes of cDNAs vary and sizes of transcripts are not as long as chromosomes. Therefore, the *N50* measurement is not as good a measure for transcriptome assemblies. It is important to assess the assembly quality by identifying full-length cDNAs [60]. Another measure can be the alignment of transcripts to reference databases [185] such as RefSeq [130] or Ensembl [50, 72].

The transcriptome assembly of the Venus flytrap (Manuscript IV) was analyzed with TargetIdentifier [112] and the assembled transcripts were aligned to RefSeq sequences of other members of the plant kingdom.

¹<https://www.broad.harvard.edu/crd/wiki/index.php/N50>, accessed 6. November 2013.

²http://korflab.ucdavis.edu/datasets/Assemblathon/Assemblathon2/Basic_metrics/assemblathon_stats.pl, accessed 16. November 2013.

Chapter 6

Gene finding and *ab initio* prediction

In computational biology gene finding or gene prediction refers to the process of assigning coding regions to genomic sequences. This process is crucial for analyses of genomes and metagenomes as it provides access to the encoded genomic information for further processing.

Conventional *ab initio* gene finding algorithms employ probabilistic algorithms describing genomic sequences containing protein and noncoding regions. Gene prediction accuracy critically depends on precision of the estimation of model parameters that are genome specific [192].

Most gene prediction algorithms are based on Hidden Markov Models. However, other approaches are available, such as dynamic programming [73] or Support Vector Machine approaches [148, 149]. One might also choose to identify genes by aligning the sequence to a database.

Various tools are available for *ab initio* gene prediction and evidence based gene predictions. A short overview of popular gene finder tools is shown in Table 6.1.

Hidden Markov models and Dynamic Programming in gene prediction

Hidden Markov models (HMMs) – proposed in the late 1960s [13] – are statistical models which can be considered to be the simplest dynamic Bayesian networks [40]. HMM models are applicable on genomic sequences due to their intrinsic order, i.e. ordered string of nucleotides. The model assumes that future states are independent of the past under the present state [40]. In other words, the present state only depends on the previous state and the probability of moving from the previous state to the present one. A

Table 6.1. Selection of gene finding tools

Tool	Target	Reference
AUGUSTUS	Eukaryote gene predictor	[157, 158]
GeneMark	Prokaryotes and eukaryotes	[192]
GLIMMER	finding genes in microbial DNA	[35]
MetaGeneMark	Metagenome gene finder	[192]
mGene	Eukaryote gene predictor	[148, 149]
Prodigal	prokaryotic gene finding	[73]

trained HMM tool can answer the question of what is the most probable path generated for a given sequence uncovering "hidden states" which could be coding/noncoding regions, splicing sites (intron/exon regions) and more [40]. Simply speaking, the model describes how likely is this sequence to be comprised of a coding or noncoding region. The assignment is dependent on a model-specific intrinsic threshold [40]. Gene finding programs such as GLIMMER [35], MetaGeneMark [192] and AUGUSTUS [157, 158] are based on the HMM principle.

Dynamic Programming (DP) – formalized in the early 1950s – is not a standardized model as HMMs. It is a term for solving complex problems by breaking them down into simpler subproblems [39]. In gene prediction, this can be done by acquiring bitwise information about the sequence, e.g. *hexamer* statistics [73]. The statistics are then employed to predict if a given sequence fulfills open reading frame requirements. Prodigal applies DP for gene prediction [73].

Sequence alignment based gene finding

Alignments based gene assignment does not fall into the typical gene finding or prediction category. This approach comprises a simple sequence alignment with alignment tools such as BLAST (or translated BLAST) [4] or UBLAST [42] to databases, e.g. RefSeq [130], ENSEMBL [50, 72] or other public databases (NCBI [142, 178] or SwissProt/TrEMBL [18]). The hit sequence is used as a template and the coding region is inferred.

6.1 Gene finding in metagenomic datasets

Sequences encoding mostly undetected homologs are abundant in new metagenomic dataset since the majority of organisms' in the environment are uncultured (described in Chapter 2). Therefore, tools which are able to identify genes with low similarity to existing database sequences are important for metagenomic studies [192]. Metagenomic gene finders are designed to fulfill this task. In the analysis of the metagenomes of the polar marine environment (Manuscript II and Manuscript III) this method was helpful in

detecting such "undetectable" proteins. The metagenomes were scanned with the two gene finding methods, (1) Prodigal (dynamic programming based [73]) and (2) MetaGeneMark (HMM based [192]) to increase the range of predicted sequences as it has been shown that MetaGeneMark had a higher precision and Prodigal a higher recall rate when calling genes in metagenomic datasets¹.

6.2 Gene finding in eukaryotes

While analyzing single genomes it is important to assign coding regions as specific for the organism's clade as possible, because for example plant genomes are different from animal genomes. This can be achieved with evidence based gene finding approaches [102, 123, 157, 158]. These methods incorporate prior knowledge about the organism such as known proteins, full-length cDNAs or expressed sequence tags (ESTs) [102]. Thus, the AUGUSTUS [157, 158] tool was used for identifying genes in the donkey genome (Manuscript I).

¹<http://genome.jgi.doe.gov/programs/metagenomes/benchmarks.jsf>, accessed 3. November 2013.

Chapter 7

From Sequence to Function and Taxonomy

7.1 Functional Annotation of Coding Regions

Gene annotation is the process of associating biological information to a sequence. It marks the next step after identifying coding regions in genomes and metagenomes. Making sense out of coding regions in single genome data, metagenomic data and RNA-seq studies gives access to the functional space of a single organism or a metagenomic community. Functional descriptions make it possible to set the genetic composition into a bigger perspective where genes and pathways can be linked to the respective environmental traits [95, 152, 167].

Annotation schemes and databases

Several annotation schemes have been proposed and are extensively used. An overview of the most used annotation designs is shown in Table 7.1. The most commonly used annotation scheme is Gene Ontology (GO) [63] for describing genes in functional categories. It provides a hierarchical, controlled vocabulary of terms that can be used to annotate gene products at varying levels of specificity [63]. The vocabulary is defined in three ontologies: molecular function, biological process, and cellular component. A gene product may be a component of one or more parts of a cell or part of the extracellular environment. A condensed vocabulary of high-level GO terms (GO Slim) can be applied to replace specific GO terms with a limited number of general-purpose ancestor terms [22].

Orthologous Group (OG) annotation is another annotation scheme. OGs

were derived by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages [161]. Each OG consists of individual proteins or groups of paralogs from at least three lineages and thus corresponds to an ancient conserved domain [161]. eggNOG [115] is one of the databases where OGs can be accessed. It includes subgroups, such as COG (Clusters of Orthologous Groups) for bacterial annotation and KOG for eukaryotic annotation [162] and non-supervised orthologous groups (NOGs) [115].

The KEGG annotation scheme on the other hand focuses on pathway descriptions [82]. It consists of pathway maps, which are collections of diagrams representing the information of pathways of interacting molecules or genes [82]. KEGG contains all known metabolic pathways and a limited, but increasing, number of regulatory pathways and molecular assemblies.

Pfam annotates proteins by domains, thus it is useful to view a protein's domain architecture. It is a collection of multiple-sequence alignments and Hidden Markov models (explained later in the chapter) of common protein domains and families [155].

Most annotation schemes are accessible through databases from which the annotation can be inferred. The UniProt database provides cross-references to most of the mentioned annotation schemes, making it an extensive resource for functional annotation [8].

Table 7.1. Commonly used functional sequence annotation schemes

Annotation	Name	Reference
GO	Gene Ontology	[63]
COG	Clusters of Orthologous Groups	[162]
NOG	Non-supervised Orthologous Groups	[115]
KEGG	Kyoto Encyclopedia of Genes and Genomes	[82]
Pfam	Protein families	[12]

Alignment based annotation

As mentioned above, the function of a gene or a protein (translated genes) can be inferred from a database. Most commonly, the sequence is aligned to the database with BLAST [4] or other alignment tools (e.g. the fast BLAST alternative UBLAST [42]). The thresholds are commonly set to an E-value cutoff of 10^{-5} and/or 50% alignment similarity over a minimum of 50% of the sequence length ("50/50 rule" [74]). This method was used in Manuscript I, to infer GO annotation to genes of the donkey genome from the UniProt database. Figure 7.1 illustrates the most abundant GO annotation categories in the assembled genome.

It can be seen that the annotation descriptions vary from detailed to generic, e.g. membrane or nucleus. Many GO terms can only be assigned

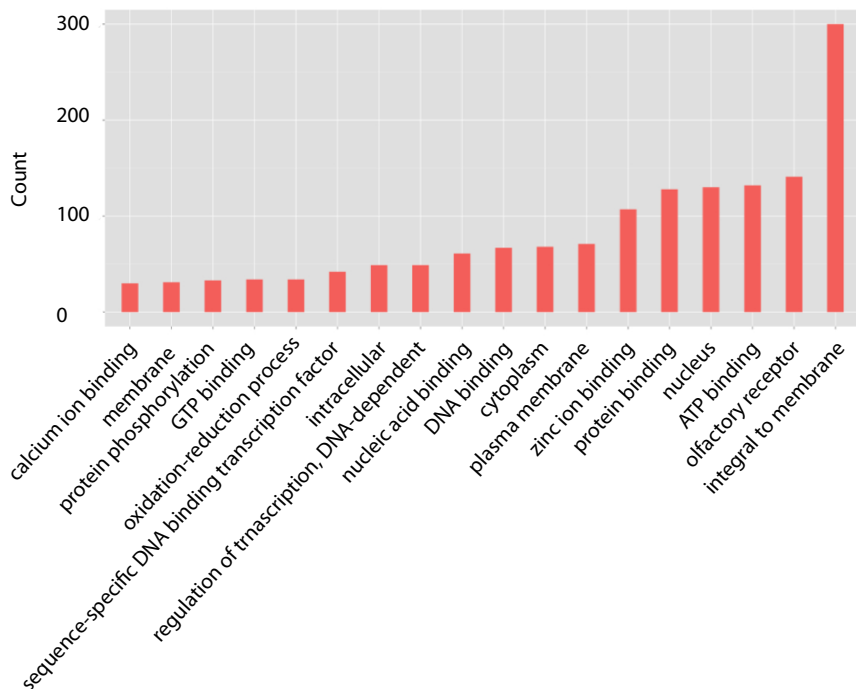


Figure 7.1. Distribution of the most abundant GO annotation categories in the donkey *de novo* assembly. Adapted from Manuscript I.

to a high level node in the GO hierarchy¹ which makes it difficult to find specific functional targets when analyzing mixed genome samples.

An alignment based annotation was also used to annotate genes from the polar marine metagenomic samples (Manuscript II and III). The annotation, however, was inferred by aligning the genes to the eggNOG database [115] for functional annotation. Based on the experience of annotating the Donkey genome, the COG and NOG annotation scheme is more extensive and less generic than the GO annotation scheme. Furthermore, the translated genes were aligned to the Swiss-Prot database via BLASTP with a 90% coverage. The EC numbers and MEROPS flags were inferred from the best hit.

Hidden Markov model based annotation

Hidden Markov model based annotation is a more directed annotation approach as it takes the active site positions into account [48, 49]. Thus,

¹<http://www.geneontology.org/GO.ontology.structure.shtml>, accessed 2. November 2013.

a single sequence can be assigned to multiple annotations. An alignment based approach, e.g. assigning GO annotations from UniProt, might not pick up a multi-domain sequence as the database annotation is not sufficient. HMM based annotations can be very specific and customizable. The Pfam annotation is based on HMMs [12]. The HMMs are based on Protein families for which individual alignments and models are precomputed [12]. An approach similar to the Pfam HMMs was used in Manuscript III for creating protease specific models to scan the polar marine metagenomes for novel sequences. The HMMs were created for protease specificity. A more elaborate description of the HMM construction is provided in Chapter 8.3.

Functional annotation of sequences can also be done by integrative software, such as InterProScan [188]. InterProScan combines several protein signature recognition methods into one resource. With this method one can scan multiple databases or HMMs for various annotations, such as Pfam and GO annotations. However, to run InterProScan on millions of genes is computationally quite expensive. InterProScan was used to identify GO annotations in Manuscript IV.

7.2 Taxonomic Annotation of Metagenomes

Metagenomic samples contain a mixture of multiple organisms (explained in Chapter 2) and their analysis mostly addresses the question of "Who is in there?". In order to find the composition of organisms in the metagenome, the processed reads (Pre-processing of sequencing raw reads was described in Chapter 2) are commonly mapped (with e.g. BWA [97] or Bowtie [92]) to target databases.

Taxonomic annotation of the metagenomes in Manuscript I and III was done by mapping to multiple databases one at a time. After mapping to the first database, unmapped reads are mapped to the next database. The reads were mapped to the following databases from top to bottom¹:

- 1st Microbial complete genomes
- 2nd Microbial draft genomes
- 3rd Viral complete genomes
- 4th Fungal complete genomes
- 5th Nucleotide database

The hit to a genome strongly depends on the mapping parameters and the mapping tool. Furthermore, a strain or species level annotation can rarely be achieved with this approach as reads can map to different genomes with the same mapping score.

In the analyses for Manuscript I and III, the lowest common taxonomy was assigned to give an overview of the metagenomic community. Therefore, representation of the environments' diversity was kept at a phylum or order

¹all databases can be accessed through <http://www.ncbi.nlm.nih.gov/>

level. The composition was represented as pie-charts, such as the taxonomic assignment of reads from the Middle Pleistocene horse sample from Thistle Creek which was analyzed in Manuscript I (Figure 7.2) [120].

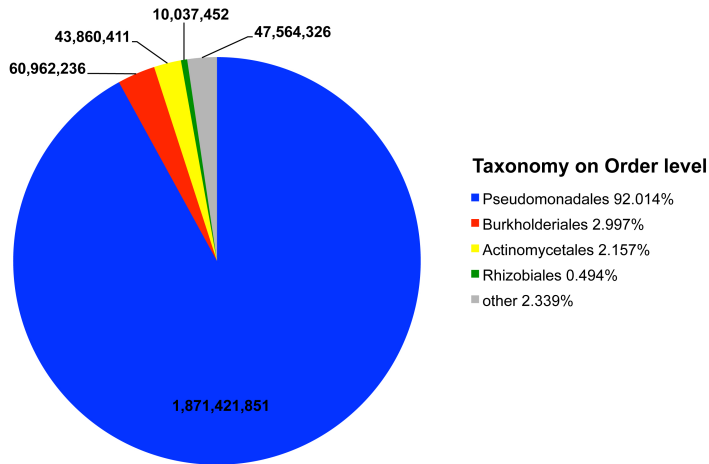


Figure 7.2. This is an example of taxonomic assignment of sequencing reads. The pie chart shows numbers of reads positively mapped to bacterial order groups. The taxonomy assignment revealed that over 92% of the reads from the old Middle Pleistocene horse sample from Thistle Creek belong to the order of Pseudomonadales. Illustration adapted from Manuscript I.

7.3 Manuscript I

In 2003, a metapodial horse sample was recovered at the Thistle Creek site in west-central Yukon Territory, Canada. The sample was dated to be approximately 560 – 780 thousand years old. This study represents the oldest full genome sequence determined so far by almost an order of magnitude. For comparison, the genome of a Late Pleistocene horse (43 kyr BP), and modern genomes of five domestic horse breeds (*Equus ferus caballus*), a Przewalski's horse (*E. f. przewalskii*) and a donkey (*E. asinus*) were sequenced. It was suggested that the Equus lineage gave rise to all contemporary horses, zebras and donkeys.

I was responsible for initial mapping of Illumina and Helicos reads and the taxonomic annotation of the Middle Pleistocene horse sample metagenome from Thistle Creek. Furthermore, I provided *ab initio* prediction of the donkey genome and the *de novo* detection of Y-chromosome scaffolds in the assembled donkey genome.

The extensive 200 page long supplement was not included in the thesis. The complete supplementary information can be accessed through nature publishing group¹. Supplementary section 4 describes the analyses of the Middle Pleistocene horse sample metagenome and the donkey genome annotation.

¹<http://www.nature.com/nature/journal/v499/n7456/full/nature12323.html#supplementary-information>

Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse

Ludovic Orlando^{1*}, Aurélien Ginolhac^{1*}, Guojie Zhang^{2*}, Duane Froese³, Anders Albrechtsen⁴, Mathias Stiller⁵, Mikkel Schubert¹, Enrico Cappellini¹, Bent Petersen⁶, Ida Moltke^{4,7}, Philip L. F. Johnson⁸, Matteo Fumagalli⁹, Julia T. Vilstrup¹, Maanasa Raghavan¹, Thorfinn Kornelissen¹, Anna-Sapfo Malaspinas¹, Josef Vogt⁶, Damian Szklarczyk^{10,†}, Christian D. Kelstrup¹⁰, Jakob Vinther^{11,†}, Andrei Dolocan¹², Jesper Stenderup¹, Amhed M. V. Velazquez¹, James Cahill⁵, Morten Rasmussen¹, Xiaoli Wang², Jiumeng Min², Grant D. Zazula¹³, Andaine Seguin-Orlando^{1,14}, Cecilie Mortensen^{1,14}, Kim Magnussen^{1,14}, John F. Thompson¹⁵, Jacobo Weinstock¹⁶, Kristian Gregersen^{1,17}, Knut H. Røed¹⁸, Vera Eisenmann¹⁹, Carl J. Rubin²⁰, Donald C. Miller²¹, Douglas F. Antczak²¹, Mads F. Bertelsen²², Søren Brunak^{6,23}, Khaled A. S. Al-Rasheid²⁴, Oliver Ryder²⁵, Leif Andersson²⁰, John Mundy²⁶, Anders Krogh^{1,4}, M. Thomas P. Gilbert¹, Kurt Kjær¹, Thomas Sicheritz-Ponten^{6,23}, Lars Juhl Jensen¹⁰, Jesper V. Olsen¹⁰, Michael Hofreiter²⁷, Rasmus Nielsen²⁸, Beth Shapiro⁵, Jun Wang^{2,26,29,30} & Eske Willerslev¹

The rich fossil record of equids has made them a model for evolutionary processes¹. Here we present a 1.12-times coverage draft genome from a horse bone recovered from permafrost dated to approximately 560–780 thousand years before present (kyr BP)^{2,3}. Our data represent the oldest full genome sequence determined so far by almost an order of magnitude. For comparison, we sequenced the genome of a Late Pleistocene horse (43 kyr BP), and modern genomes of five domestic horse breeds (*Equus ferus caballus*), a Przewalski's horse (*E. f. przewalskii*) and a donkey (*E. asinus*). Our analyses suggest that the *Equus* lineage giving rise to all contemporary horses, zebras and donkeys originated 4.0–4.5 million years before present (Myr BP), twice the conventionally accepted time to the most recent common ancestor of the genus *Equus*^{4,5}. We also find that horse population size fluctuated multiple times over the past 2 Myr, particularly during periods of severe climatic changes. We estimate that the Przewalski's and domestic horse populations diverged 38–72 kyr BP, and find no evidence of recent admixture between the domestic horse breeds and the Przewalski's horse investigated. This supports the contention that Przewalski's horses represent the last surviving wild horse population⁶. We find similar levels of genetic variation among Przewalski's and domestic populations, indicating that the former are genetically viable and worthy of conservation efforts. We also find evidence for continuous selection on the immune system and olfaction throughout horse evolution. Finally, we identify 29 genomic regions among horse breeds that deviate from neutrality and show low levels of genetic variation compared to the Przewalski's horse. Such regions could correspond to loci selected early during domestication.

In 2003, we recovered a metapodial horse fossil at the Thistle Creek site in west-central Yukon Territory, Canada (Fig. 1a). The fossil was

from an interglacial organic unit associated with the Gold Run volcanic ash, dated to 735 ± 88 kyr BP^{2,3} (Fig. 1b). Relict ice wedges below the unit indicate persistent permafrost since deposition (Supplementary Information, section 1.1), whereas the organic unit, hosting the fossil, indicates a period of permafrost degradation, or a thaw unconformity⁷, during a past interglacial as warm or warmer than present³, and rapid deposition during either marine isotope stage 19, 17 or 15. This indicates that the fossil dates to approximately 560–780 kyr BP. The metapodial shows typical caballine morphology, consistent with Middle rather than the smaller Late Pleistocene horse fossils from the area (Fig. 1c and Supplementary Information, section 1.2). This age is consistent with small mammal fossils from this unit indicating a Late Irvingtonian, or Middle Pleistocene, age³, and infinite radiocarbon dates⁸.

Theoretical⁹ and empirical evidence¹⁰ indicates that this age approaches the upper limit of DNA survival. So far, no genome-wide information has been obtained from fossil remains older than 110–130 kyr BP¹¹. Time-of-flight secondary ion mass spectrometry (TOF-SIMS) on the ancient horse bone revealed secondary ion signatures typical of collagen within the bone matrix (Fig. 2a and Supplementary Table 7.1), and high-resolution tandem mass spectrometry sequencing¹² revealed 73 proteins, including blood-derived peptides (Supplementary Information, section 7.4). This is consistent with good biomolecular preservation, suggesting possible DNA survival. Therefore, we conducted larger-scale destructive sampling for genome sequencing.

We used Illumina and Helicos sequencing to generate 12.2 billion DNA reads from the Thistle Creek metapodial. Mapping against the horse reference genome yielded $\sim 1.12\times$ genome coverage. We based the size distribution of ancient DNA templates on collapsed Illumina

¹Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark. ²Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen 518083, China. ³Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta T6G 2E3, Canada. ⁴The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ⁵Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064, USA. ⁶Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. ⁷Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA. ⁸Department of Biology, Emory University, Atlanta, Georgia 30322, USA. ⁹Department of Integrative Biology, University of California, Berkeley, California 94720, USA. ¹⁰Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark. ¹¹Jackson School of Geosciences, The University of Texas at Austin, 1 University Road, Austin, Texas 78712, USA. ¹²Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78712, USA. ¹³Government of Yukon, Department of Tourism and Culture, Yukon Palaeontology Program, PO Box 2703 L2A, Whitehorse, Yukon Territory Y1A 2C6, Canada. ¹⁴Danish National High-throughput DNA Sequencing Centre, University of Copenhagen, Øster Farimagsgade 2D, 1353 Copenhagen K, Denmark. ¹⁵NABSys Inc, 60 Clifford Street, Providence, Rhode Island 02903, USA. ¹⁶Archeology, University of Southampton, Avenue Campus, Highfield, Southampton SO17 1BF, UK. ¹⁷Zoological Museum, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark. ¹⁸Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, Box 8146 Dep, N-0033 Oslo, Norway. ¹⁹Département histoire de la Terre, UMR 5143 du CNRS, paléobiodiversité et paléoenvironnements, MNHN, CP 38, 8, rue Buffon, 75005 Paris, France. ²⁰Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden. ²¹Baker Institute for Animal Health, Cornell University, Ithaca, New York 14853, USA. ²²Center for Zoo and Wild Animal Health, Copenhagen Zoo, 2000 Frederiksberg, Denmark. ²³Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark. ²⁴Zoology Department, College of Science, King Saud University, Riyadh 11451, Saudi Arabia. ²⁵San Diego Zoo's Institute for Conservation Research, Escondido, California 92027, USA. ²⁶Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ²⁷Department of Biology, The University of York, Wentworth Way, Heslington, York YO10 5DD, UK. ²⁸Departments of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, California 94720, USA. ²⁹King Abdulaziz University, Jeddah 21589, Saudi Arabia. ³⁰Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [†]Present addresses: Bioinformatics Group, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland (D.S.); Departments of Earth Sciences and Biological Sciences, University of Bristol BS8 1UG, UK (Ja.V.).

*These authors contributed equally to this work.

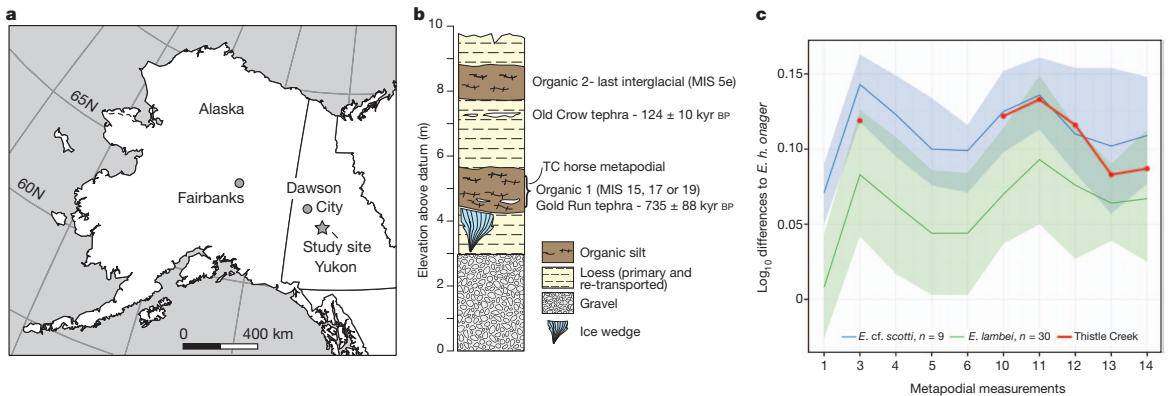


Figure 1 | The early Middle Pleistocene horse metapodial from Thistle Creek (TC). **a**, Geographical localization. **b**, Stratigraphic setting. **c**, Morphological comparison to Middle and Late Pleistocene horses from Beringia. Simpson's ratio diagrams contrasting \log_{10} differences in 10 metapodial measurements between horse fossils and a reference (*E. hemionus onager*) are shown for a series of 9 and 30 horses from the Middle and the Late Pleistocene era, respectively (Supplementary Information, section 1.2). The full

read pairs (Supplementary Fig. 4.4), yielding an average length of 77.5 base pairs (bp). The specimen is male based on X to autosomal chromosome coverage (Supplementary Information, section 4.2b) and the presence of Y-chromosome markers (Supplementary Information, section 4.1d). Endogenous read content was lower for Illumina (0.47%) than Helicos (4.21%) using standard⁸ or improved¹³ single-strand template preparation procedures. This is probably due to 3' ends available at nicks, resistance of undamaged modern DNA contaminants to denaturation, and Helicos ability to sequence short templates. Despite this, endogenous DNA content was >16.6–20.0-fold lower than for Saqqaq Palaeo-Eskimo¹⁴ and Denisovan specimens¹⁵, both sequenced to high depth.

Several observations support genome sequence authenticity. First, a 348-bp mitochondrial control region segment was replicated independently (Supplementary Fig. 2.2 and Supplementary Information, section 2.4). Second, phylogenetic analyses on data obtained with two sequencing platforms in different laboratories are consistent (Supplementary Fig. 8.4), ruling out post-purification contamination. Third, autosomal, Y-chromosomal and mitochondrial DNA analyses place the Thistle Creek specimen basal to Late Pleistocene and modern horses (Fig. 3a and Supplementary Figs 8.1–8.4). Fourth, we found signs of severe biomolecular degradation, including levels of cytosine deamination at overhangs considerably higher than observed in 28 younger permafrost-preserved fossils from the Late Pleistocene (Fig. 2c, Supplementary Fig. 6.40 and Supplementary Table 6.1) and protein deamidation levels^{12,16} (Fig. 2b and Supplementary Information, section 7.5) greater than those reported for younger permafrost-preserved bones.

We additionally sequenced genomes of a 43-kyr-old (pre-domestication) horse (1.8× coverage), a modern donkey (16×; Supplementary Fig. 4.1), 5 modern domestic horses (Arabian, Icelandic, Norwegian fjord, Standardbred and Thoroughbred; 7.9×–21.1×) and one modern Przewalski's horse (9.6×; Supplementary Table 2.1), considered to possibly represent the last surviving wild horse population. We used this data set to address fundamental questions in horse evolution: (1) the timing of the origins of the genus *Equus*; (2) the demographic history of modern horses; (3) the divergence time of horse populations forming the Przewalski's and domestic lineages; (4) the extent to which the Przewalski's horse has remained isolated from domestic relatives; (5) the timing of gene expansions within the horse genome; (6) the identification of genes potentially under selection during horse evolution.

As no accepted *Equus* fossils exist before 2.0 Myr BP^{4,5} (Supplementary Information, section 9.1d), the date of the last common ancestor that

distribution range between minimal and maximal values is presented within shaded areas. Numbers reported on the x axis refer to the following measurements: 1, maximal length; 3, breadth at the middle of the diaphysis; 4, depth at the middle of the diaphysis; 5, proximal breadth; 6, proximal depth; 10, distal supra-articular breadth; 11, distal articular breadth; 12, depth of the keel; 13, least depth of medial condyle; 14, greatest depth of medial condyle.

gave rise to extant horses versus donkeys, asses and zebras¹⁷ remains heavily debated. Proposed dates extend as early as 4.2–4.5 Myr BP on the basis of palaeontological estimates¹⁸ to over 6.0 Myr BP according to molecular analyses¹⁹. We addressed this issue by taking advantage of the established age for the Thistle Creek horse. As a sample cannot be older than the population it belonged to, we explored a full range of possible calibrations for the *Equus* most recent common ancestor (MRCA) and calculated the divergence time between the populations of the ancient Thistle Creek horse and modern horses²⁰ (Supplementary Information, section 10.1). Calibrations resulting in divergence times younger than the Thistle Creek bone age were rejected, providing a credible confidence range for the MRCA of *Equus*. We found rates consistent with the *Equus* MRCA living 3.6–5.8 Myr BP to be compatible with our data (Fig. 3b and Supplementary Figs 10.1–10.3). We also found support for slower mutation rates in horse than human (Supplementary Information, section 8.4 and Supplementary Table 8.5), implying a minimal date of 4.07 Myr BP for the MRCA of *Equus* (Supplementary Figs 10.1–10.3). We therefore propose 4.0–4.5 Myr BP for the MRCA of all living *Equus*, in agreement with recent molecular findings¹⁷ and the oldest palaeontological records for the monodactyle *Plesippus simplicidens*, which some¹⁸ consider the earliest fossil of *Equus*. Our result indicates that the evolutionary timescale for the origin of contemporary equid diversity is at least twice that commonly accepted.

Second, we reconstructed horse population demography over the last 2 Myr. The pairwise sequential Markovian coalescent (PSMC) approach²¹ shows that horses experienced a population minimum approximately 125 kyr BP, corresponding to the last interglacial when environmental conditions were similar to now throughout their range. The population expanded during the cold stages of marine isotope stage (MIS) 4 and 3 as grasslands expanded. A peak was reached 25–50 kyr BP and was followed by an approximately 100-fold collapse, probably resulting from major climatic changes and related grassland contraction after the Last Glacial Maximum²² (Fig. 4 and Supplementary Figs 9.4–9.5). A similar demographic history was inferred from Bayesian skyline reconstructions using 23 newly characterized ancient mitochondrial genomes (Supplementary Fig. 9.6). These results support suggestions²² that climatic changes are major demographic drivers for horse populations. PSMC analyses also revealed two earlier demographic phases (Fig. 4b and Supplementary Figs 9.4–9.5), with population sizes peaking 190–260 kyr BP and 1.2–1.6 Myr BP, respectively, followed by 1.7-fold and 8.1-fold collapses. Extremely low population sizes were inferred approximately 500–800 kyr BP, a time period

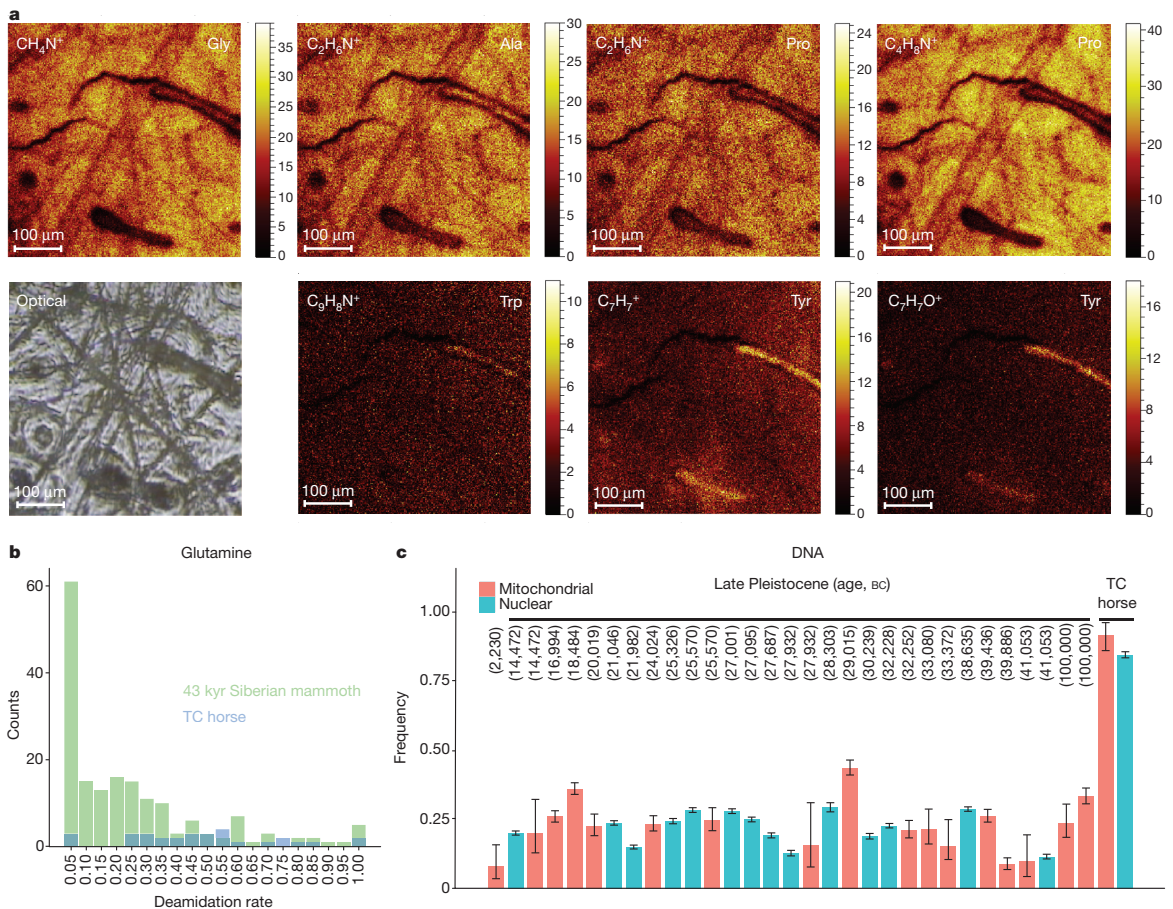


Figure 2 | Amino acid, protein and DNA preservation of the Thistle Creek horse bone. **a**, Amino acid signatures. Secondary ions, characteristic of five amino acids over- or under-represented in collagen, were detected by TOF-SIMS (Supplementary Information, section 7.1). The size of secondary ion maps is $500 \times 500 \mu\text{m}^2$ with a resolution of 256×256 pixels. **b**, Glutamine deamidation. The observed distribution of glutamine deamidation levels (Supplementary Information, section 7.5) is blue for the Thistle Creek (TC) horse bone and green for a 43-kyr-old Siberian mammoth bone.

that covers the divergence time of the Thistle Creek and contemporary horse populations. This result may relate to population fragmentation when horses colonized Eurasia from America, in agreement with the earliest presence of horses in Eurasia 750 kyr BP⁴.

We next investigated whether Przewalski's horse indeed represents the last survivor of wild horses. Native to the Mongolian steppes, this horse was listed as extinct in the wild (IUCN red list²³) but has been reassigned to endangered after successful conservation and reintroduction. Using maximum likelihood phylogenetic analyses and topological tests (Supplementary Information, sections 8.2–8.3), we found that the Przewalski's horse genome falls outside a monophyletic group of domestic horses. The MRCA of Przewalski's and domestic horse sequences dates to 341–431 kyr BP (Supplementary Table 8.3), a period consistent with previous estimates⁶. We estimated the divergence time between populations of Przewalski's and domestic horses to approximately 38–72 kyr BP (Supplementary Tables 10.4–10.6). Our 43 kyr BP horse genome branched off before the Przewalski's and domestic horse lineages diverged (Fig. 3a). This specimen belonged to a population that diverged from that leading to modern horses approximately 89–167 kyr BP

(Supplementary Figs 10.1–10.3 and Supplementary Table 10.5), providing a maximal boundary for the younger divergence between Przewalski's and domestic horses.

Using quartet alignments and *D* statistics²⁴ (Supplementary Information, sections 12.1–12.3) we found no evidence for admixture between the Przewalski's horse and the individual horse breeds investigated in this study using either the donkey or the ancient Thistle Creek genome as out-group (Supplementary Tables 12.1–S12.3). Scanning the Przewalski's horse genome, we also found no long tracts of shared polymorphisms with domestic horses (Supplementary Fig. 12.3), as would be expected if recent admixture occurred after the last wild individual was captured in the 1940s²⁵. Rather, we identified long tracts of variation unique to the Przewalski's horse genome, including genes involved in immunity, cytoskeleton, metabolism and the central nervous system that could have been specifically selected in this lineage (Supplementary Information, section 12.6). The average levels of polymorphism present in the Przewalski's horse genome are greater than those observed in the Icelandic, Standardbred and Arabian horse genomes (Supplementary Fig. 5.5 and Supplementary Table 11.10). Thus, unadmixed lineages

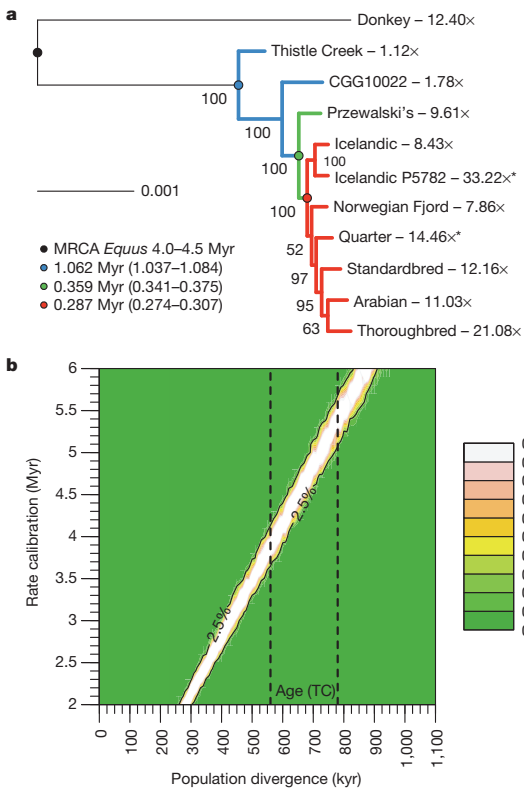


Figure 3 | Horse phylogenetic relationships and population divergence times. **a**, Maximum likelihood phylogenetic inference. We performed a supermatrix analysis of 5,359 coding genes (Supplementary Information, section 8.3a, 100 bootstrap pseudo-replicates) and estimated the average age for the main nodes (r8s semi-parametric penalized likelihood (PL) method, Supplementary Information, section 8.3c; see Supplementary Table 8.3 for other analyses). Asterisk indicates previously published horse genomes. **b**, Population divergence times. We used ABC to recover a posterior distribution for the time when two horse populations split over a full range of possible mutation rate calibrations (Supplementary Information, section 10.1). The first population included the Thistle Creek horse; the second consisted of modern domestic horses. A conservative age range for the Thistle Creek horse is reported between the dashed lines (560–780 kyr).

for certain functionally important gene families²⁶ (Supplementary Information, section 5.1c). Our data set revealed that a limited fraction of horse paralogues (1.7%, representing 258 paralogues) showed no hits among donkey reads, suggesting that most horse paralogues expanded before the origin of the genus *Equus* some 4.0–4.5 Myr BP. Among these 258 regions, 11 L1 retrotransposons and one copy of a keratin gene are absent from the ancient Thistle Creek horse genome but present in the 43 kyr horse and modern horses (Supplementary Table 5.3), suggesting an expansion before their MRCA some 500–626 kyr BP (Supplementary Table 8.3). Similarly, 44 L1-retrotransposon paralogues were found only in modern horse genomes (Supplementary Table 5.4), indicating that expansion of L1 retrotransposons has remained active since then.

Finally, we identified loci potentially selected in modern horses (Supplementary Figs 11.1–11.2), focusing on regions showing unusual densities of derived mutations (Supplementary Information, section 11.1). We caution that local variations in mutation and recombination rates, as well as misalignments, may result in similar signatures at neutrally evolving regions. Functional clustering analyses revealed significant enrichment for immunity-related and olfactory receptor genes (Supplementary Table 11.4), two categories also enriched for non-synonymous single nucleotide polymorphisms (SNPs) (Supplementary Information, section 5.2d). Additionally, we identified 29 regions showing deviation from neutrality and significant reduction in genetic diversity among modern domestic horses compared to Przewalski's horse (Supplementary Tables 11.8–11.9). Such regions could correspond to loci that have been selected and transmitted to all horse breeds investigated here after divergence from the Przewalski's horse population,

are still present in the endangered Przewalski's horse population, with levels of allelic diversity that can support long-term survival of captive breeding stocks despite descending from only 13–14 wild individuals²⁵.

The sequencing of the horse reference genome showed increased paralogous expansion rates in horses compared to humans and bovines

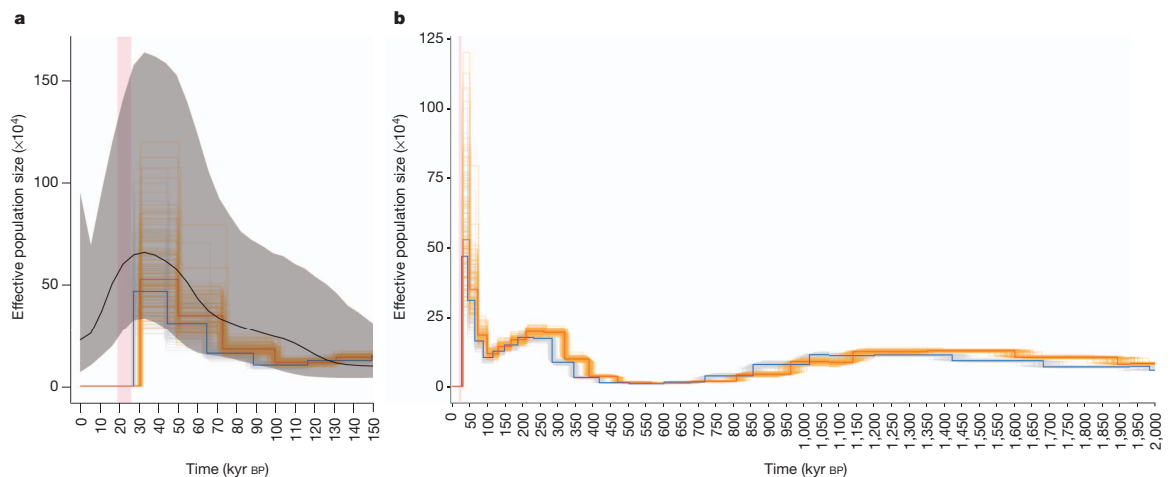


Figure 4 | Horse demographic history. **a**, Last 150 kyr BP. PSMC based on nuclear data (100 bootstrap pseudo-replicates) and Bayesian skyline inference based on mitochondrial genomes (median, black; 2.5% and 97.5% quantiles, grey) are presented following the methodology described in Supplementary Information, section 9. The Last Glacial Maximum (19–26 kyr BP) is shown in

pink. **b**, Last 2 Myr BP. PSMC profiles are scaled using the new calibration values proposed for the MRCA of all living members of the genus *Equus* (4.0 Myr, blue; 4.5 Myr, red), and assuming a generation time of 8 years (for other generation times, see Supplementary Figs 9.4 and 9.5).

possibly related to domestication. These regions include genes for the KIT ligand critical for haematopoiesis, spermatogenesis and melanogenesis, and myopalladin involved in sarcomere organization.

Our study has pushed the timeframe of palaeogenomics back by almost an order of magnitude. This enabled us to readdress a range of questions related to the evolution of *Equus*—a group representing textbook examples of evolutionary processes. The Thistle Creek genome also provided us with direct estimates of the long-term rate of DNA decay²⁷, revealing that a significant fraction (6.0–13.3%) of short (25-bp) DNA fragments may survive over a million years in the geosphere (Supplementary Fig. 6.42). Thus, procedures maximizing the retrieval of short, but still informative, DNA may provide access to resources previously considered to be much too old. Methods have recently been developed for increasing the sequencing depth of ancient genomes¹⁵ but do not increase the percentage of endogenous sequences retrieved. Overcoming this technical challenge with whole-genome enrichment approaches, and lower sequencing costs, will make retrieval of higher coverage genomes from specimens with low endogenous DNA content practical and economical.

METHODS SUMMARY

Ancient horse extracts and DNA libraries were prepared in facilities designed to analyse ancient DNA following standard procedures^{8,12}. Protein sequencing was performed using nanoflow liquid chromatography tandem mass spectrometry²⁸. DNA sequencing was performed using Illumina and Helicos sequencing platforms^{8,13}. Reads were aligned to the horse reference genome²⁶ and *de novo* assembled donkey scaffolds using BWA²⁹. Maximum-likelihood DNA damage rates were estimated from nucleotide misincorporation patterns. Population divergence times were estimated disregarding transitions to limit the impact of replication of damaged DNA and following ref. 20 with quartet genome alignments instead of trios and implementing approximate Bayesian computation (ABC).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 October 2012; accepted 30 May 2013.

Published online 26 June 2013.

- Franzen, J. L. *The Rise of Horses: 55 Million Years of Evolution* (Johns Hopkins Univ. Press, 2010).
- Froese, D. G., Westgate, J. A., Reyes, A. V., Enkin, R. J. & Preece, S. J. Ancient permafrost and a future, warmer Arctic. *Science* **321**, 1648 (2008).
- Westgate, J. A. *et al.* Gold Run tephra: A Middle Pleistocene stratigraphic and paleoenvironmental marker across west-central Yukon Territory, Canada. *Can. J. Earth Sci.* **46**, 465–478 (2009).
- Eisenmann, V. Origins, dispersals, and migrations of *Equus* (Mammalia, Perissodactyla). *Courier Forschungsintitut Senckenberg* **153**, 161–170 (1992).
- Forsten, A. Mitochondrial-DNA timetable and the evolution of *Equus*: Comparison of molecular and paleontological evidence. *Ann. Zool. Fenn.* **28**, 301–309 (1992).
- Goto, H. *et al.* A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. *Genome Biol. Evol.* **3**, 1096–1106 (2011).
- Reyes, A. V., Froese, D. G. & Jensen, B. J. Response of permafrost to last interglacial warming: field evidence from non-glaciated Yukon and Alaska. *Quat. Sci. Rev.* **29**, 3256–3274 (2010).
- Orlando, L. *et al.* True single-molecule DNA sequencing of a Pleistocene horse bone. *Genet. Res.* **21**, 1705–1719 (2011).
- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317**, 111–114 (2007).
- Miller, W. *et al.* Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl Acad. Sci. USA* **109**, E2382–E2390 (2012).
- Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J. Proteome Res.* **11**, 917–926 (2012).
- Ginolhac, A. *et al.* Improving the performance of True Single Molecule Sequencing for ancient DNA. *BMC Genomics* **13**, 177 (2012).
- Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
- Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
- van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid Commun. Mass Spectrom.* **26**, 2319–2327 (2012).
- Vilstrup, J. T. *et al.* Mitochondrial phylogenomics of modern and ancient equids. *PLoS ONE* **8**, e55950 (2013).
- McFadden, B. J. & Carranza-Castaneda, O. Cranium of *Dinohippus mexicanus* (Mammalia Equidae) from the early Pliocene (latest Hemphillian) of central Mexico and the origin of *Equus*. *Bull. Florida Museum Nat. History* **43**, 163–185 (2002).
- Weinstock, J. *et al.* Evolution, systematics, and phylogeography of Pleistocene horses in the new world: a molecular perspective. *PLoS Biol.* **3**, e241 (2005).
- Green, R. E. *et al.* A draft sequence of the Neanderthal genome. *Science* **328**, 710–722 (2010).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* **479**, 359–364 (2011).
- International Union for Conservation of Nature. IUCN Red List of Threatened Species, Version 2010.1, <http://www.iucnredlist.org> (downloaded 11 March 2010).
- Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
- Bowling, A. T. *et al.* Genetic variation in Przewalski's horses, with special focus on the last wild caught mare, 231 Orizlita III. *Cytogenet. Genome Res.* **102**, 226–234 (2003).
- Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867 (2009).
- Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. Lond. B* **279**, 4724–4733 (2012).
- Kelstrup, C. D., Young, C., Lavallee, R., Nielsen, M. L. & Olsen, J. V. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *J. Proteome Res.* **11**, 3487–3497 (2012).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Brand, the laboratory technicians at the Danish National High-throughput DNA Sequencing Centre and the Illumina sequencing platform at SciLifeLab-Uppsala for technical assistance; J. Clausen for help with the donkey samples; S. Rasmussen for computational assistance; J. N. MacLeod and T. Kalbfleisch for discussions involving the re-sequencing of the horse reference genome; and S. Sawyer for providing published ancient horse data. This work was supported by the Danish Council for Independent Research, Natural Sciences (FNU); the Danish National Research Foundation; the Novo Nordisk Foundation; the Lundbeck Foundation (R52-A5062); a Marie-Curie Career Integration grant (FP7 CIG-293845); the National Science Foundation ARC-0909456; National Science Foundation DBI-0906041; the Searle Scholars Program; King Saud University Distinguished Scientist Fellowship Program (DSFF); Natural Science and Engineering Research Council of Canada; the US National Science Foundation DMR-0923096; and a grant RC2 HG005598 from the National Human Genetics Research Institute (NHGRI). A.G. was supported by a Marie-Curie Intra-European Fellowship (FP7 IEF-299176). M.F. was supported by EMBO Long-Term Post-doctoral Fellowship (ALTF 229-2011). A.-S.M. was supported by a fellowship from the Swiss National Science Foundation (SNSF). Mi.S. was supported by the Lundbeck foundation (R82-5062).

Author Contributions L.O. and E.W. initially conceived and headed the project; G.Z. and Ju.W. headed research at BGI; L.O. and E.W. designed the experimental research project set-up, with input from B.S. and R.N.; D.F. and G.D.Z. provided the Thistle Creek specimen, stratigraphic context and morphological information, with input from K.K.; K.H.R., B.S., K.G., D.C.M., D.F.A., K.A.S.A.-R. and M.F.B. provided samples; L.O., J.T.V., Ma.R., M.H., C.M. and J.S. did ancient and modern DNA extractions and constructed Illumina DNA libraries for shotgun sequencing; Ja.W. did the independent replication in Oxford; Ma.S. did ancient DNA extractions and generated target enrichment sequence data; Ji.M. and X.W. did Illumina libraries on donkey extracts; K.M., C.M. and A.S.-O. performed Illumina sequencing for the Middle Pleistocene and the 43-kyr-old horse genomes, the five domestic horse genomes and the Przewalski's horse genome at Copenhagen, with input from Mo.R.; Ji.M. and X.W. performed Illumina sequencing for the Middle Pleistocene and the donkey genomes at BGI; J.F.T. headed true Single DNA Molecule Sequencing of the Middle Pleistocene genome; A.G., B.P. and Mi.S. did the mapping analyses and generated genome alignments, with input from L.O. and A.K.; Jo.V. and T.S.-P. did the metagenomic analyses, with input from A.G., B.P., S.B. and L.O.; Jo.V. and T.S.-P. did the *ab initio* prediction of the donkey genes and the identification of the Y chromosome scaffolds, with input from A.G. and Mi.S.; L.O., A.G. and P.L.F.J. did the damage analyses, with input from I.M.; A.G. did the functional SNP assignment; A.M.V.V. and L.O. did the PCA analyses, with input from O.R.; B.S. did the phylogenetic and Bayesian skyline reconstructions on mitochondrial data; Mi.S. did the phylogenetic and divergence dating based on nuclear data, with input from L.O.; A.G. did the PSMC analyses using data generated by C.J.R. and L.A.; L.O. and A.G. did the population divergence analyses, with input from J.C., R.N. and M.F.; L.O., A.G. and T.K. did the selection scans, with input from A.-S.M. and R.N.; A.A., I.M. and M.F. did the admixture analyses, with input from R.N.; L.O. and A.G. did the analysis of paralogues and structural variation; Ja.V. and A.D. did the amino-acid composition analyses; E.C., C.D.K., D.S., L.J.J. and J.V.O. did the proteomic analyses, with input from M.T.P.G. and A.M.V.V.; L.O. and V.E. performed the morphological analyses, with input from D.F. and G.D.Z.; L.O. and E.W. wrote the manuscript, with critical input from M.H., B.S., Jo.M. and all remaining authors.

Author Information All sequence data have been submitted to Sequence Read Archive under accession number SRA082086 and are available for download, together with final BAM and VCF files, *de novo* donkey scaffolds, and proteomic data at <http://geogenetics.ku.dk/publications/middle-pleistocene-omics>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.O. (Lorlando@snm.ku.dk), Ju.W. (wangjun30@gmail.com) or E.W. (ewillerslev@snm.ku.dk).

METHODS

Genome sequencing. All fossil specimens were extracted in facilities designed to analyse ancient DNA using silica-based extraction procedures^{30,31} (Supplementary Information, section 2). A total number of 16 ancient horse extracts were built into Illumina libraries (Supplementary Information, section 2) and shotgun-sequenced at the Centre for GeoGenetics (Supplementary Tables 2.3 and 4.9). The full mitochondrial genome of a total number of 16 ancient horse specimens was captured using MYselect in-solution target enrichment kit (Supplementary Information, section 3.3b) following library construction³², and sequenced at Penn State/UCSC (Supplementary Tables 2.4 and 4.10). The combination of shotgun sequencing and capture-based sequencing performed in those two laboratories resulted in the characterization of 23 novel pseudo-complete ancient horse mitochondrial genomes (Supplementary Table 8.1). Additional sequencing was compatible with the characterization of draft nuclear genomes of two ancient horse specimens (Supplementary Tables 4.9 and 4.11): that of a Middle Pleistocene horse from Thistle Creek (560–780 kyr BP), and that of a Late Pleistocene horse from the Taymyr Peninsula (CGG10022, cal. 42,012–40,094 BC; Supplementary Table 2.3). The Thistle Creek horse draft genome was characterized using Illumina (11,593,288,435 reads, Supplementary Table 3.2; coverage = 0.74×, Supplementary Table 4.11) and Helicos sequence data (654,292,583 reads, Supplementary Table 3.5; coverage = 0.38 ×, Supplementary Table 4.11). Ancient specimens were radiocarbon dated at Belfast 14Chrono facilities (Supplementary Tables 2.3 and 2.4). The Middle Pleistocene Thistle Creek horse bone is associated with infinite radiocarbon dates.

Modern equine genomes from five modern horse breeds (Arabian, Icelandic, Norwegian fjord, Standardbred, Thoroughbred), one Przewalski's horse individual and one domestic donkey were characterized using Illumina paired-end sequencing (Supplementary Information, sections 3.1.b.3–3.1.b.4). DNA was extracted and prepared into libraries (Supplementary Information, section 2.2) in laboratories located in buildings physically separated from ancient DNA laboratory facilities. Modern horse genomes were sequenced at the Danish National High-Throughput DNA Sequencing Centre whereas the donkey genome was characterized at BGI, Shenzhen (Supplementary Information, 3.1). Trimmed reads were aligned to the horse reference genome EquCab2.0 (ref. 26), excluding the mitochondrial genome and chrUn, using BWA²⁹ (Supplementary Information, section 4.2). We generated a draft *de novo* assembly of the donkey genome using de Bruijn graphs as implemented within SOAPdenovo³³ (Supplementary Information, section 4.1.a), built gene models using Augustus³⁴ and SpyPhy³⁵ (Supplementary Information, section 4.1.b), and identified candidate scaffolds originating from the X and Y chromosomes (Supplementary Information, sections 4.1.c and 4.1.d). Sequence reads were also aligned against *de novo* assembled donkey scaffolds (Supplementary Information, section 4.2). For all genomes characterized in this study, we estimated that overall error rates were low (Supplementary Information, section 4.4.a), with type-specific error rates inferior to 5.3×10^{-4} , except for ancient genomes where post-mortem DNA damage inflated the GC→AT mis-incorporation rates (Supplementary Table 4.12). Metagenomic assignment of all reads generated from the Thistle Creek horse bone was performed using BWA-sw³⁶ and mapping against a customized database, which included all bacterial, fungal and viral genomes available (Supplementary Information, section 4.3).

Genomic variation. SNPs were called for modern genomes using the mpileup command from SAMtools (0.1.18)³⁷ and bcftools, and were subsequently filtered using vcftools varFilter and stringent quality filter criteria (Supplementary Information, section 5.2). We compared overall SNP variation levels (Supplementary Information, sections 5.2b and 11.2; Supplementary Table 11.10) present in modern horse genomes. We also compared genotypic information extracted from the genomes characterized in this study to that of 362 horse individuals belonging to 14 modern domestic breeds and 9 Przewalski's horses³⁸. Genotype and the breed/population of origin were converted into PLINK map and ped formats³⁹ and further analysed using the software Smartpca of EIGENSOFT 4.0 (ref. 40). PCA plots were generated using R 2.12.2 (ref. 41) (Supplementary Figs 5.6–5.14). Filtered SNPs that passed our quality criteria (Supplementary Information, section 5.2.a) were categorized into a series of functional and structural genomic classes using the Perl script variant_effect_predictor.pl version 2.5 (ref. 42) available at Ensembl and the EquCab2.0 annotation database version 65 (Supplementary Information, section 5.2b). We also screened our genome data for a list of 36 loci that have been associated with known phenotypic defects and/or variants (Supplementary Information, section 5.2e and Supplementary Tables 5.19 and 5.20). We systematically looked in the donkey genome for the presence of genes that have been identified in the horse reference genome as paralogues. This was performed by downloading from Ensembl a list of 15,310 paralogues and extracting genomic coordinates of the 15,171 paralogues that were located on the 31 autosomes and the X chromosome. We next calculated the average depth-of-coverage of these regions using the alignment of donkey reads against the horse reference genome. A total number of 258 paralogues exhibited no hit and were

putatively missing from the donkey genome. We further tested for the presence of those paralogues in the different ancient horse genomes characterized here, using a model where observed depth-of-coverage in ancient individual (Illumina data) is a function of the depth-of-coverage observed in a modern horse male individual, local %GC and read length (Supplementary Information, section 5.1.c). A similar model was used for identifying segmental duplications in modern equid genomes (Supplementary Information, section 5.1b).

DNA damage. We estimated DNA damage levels in the Thistle Creek horse sample and compared these to the DNA damage levels observed among other Pleistocene horse fossil bones, all associated with more recent ages (Supplementary Tables 2.3 and 2.4). All fossil specimens analysed were permafrost-preserved, limiting environmental-dependent variation in DNA damage rates⁴³. DNA fragmentation and nucleotide mis-incorporation patterns were plotted using the mapDamage package⁴⁴ (Supplementary Information, section 6.2). We then developed a DNA damage likelihood model after the model presented in ref. 45, with slight modifications, where ancient DNA fragments consist of four non-overlapping regions from 5' to 3' ends: (1) a single-stranded overhang; (2) a double-stranded region that extends until a single-strand break is encountered; (3) a double-stranded region that extends 3' of the single strand break previously mentioned, and; (4) a single stranded overhang (Supplementary Information, section 6.3 and Supplementary Fig. 6.39). All model parameters were estimated using maximum likelihood. Confidence intervals were found by taking each parameter in turn and slowly adjusting that parameter while maximizing the likelihood with respect to all other parameters until finding the points above and below with likelihood 1.92 units below the maximum. Finally, we used the model framework presented in ref. 27 to recover direct estimates of DNA survival rates from next-generation sequence data (Supplementary Information, section 6.4). We restricted our analyses (1) to the distribution of templates showing sizes superior to the modal size category; and (2) to collapsed paired-end reads, as the size of the latter corresponds to the exact size of ancient DNA fragments inserted in the DNA library.

Amino acid and proteomic analyses. A sample of the Middle Pleistocene Thistle Creek horse bone was embedded in Epothin resin under sterile conditions, cut and polished until chemical analysis of the sample surface could be performed with a time-of-flight secondary ion mass spectrometer (TOF-SIMS) instrument (Supplementary Information, section 7). We also performed high-resolution mass spectrometry (MS)-based shotgun proteomics analysis using two fragments from the Middle Pleistocene Thistle Creek horse bone (weighing 86 and 78 mg, respectively) in order to retrieve large-scale molecular information. The overall methodological approach follows the procedure that was previously applied to survey the remains of the bone proteome from three mammoth specimens living approximately 11–43 kyr ago¹², although with significant improvements (Supplementary Information, sections 7.2–7.3). Strict measures to avoid contamination and exclude false-positive results were implemented at every step, allowing to confidently profile 73 ancient bone proteins (from the attribution of 659 unique peptides based on 13,030 spectra). Raw spectrum files were searched on a local workstation using the MaxQuant algorithm version 1.2.2.5 (ref. 46) and the Andromeda peptide search engine⁴⁷ against the target/reverse list of horse proteins available from Ensembl (EqCab2.64.pep.all), the IPI v.3.37 human protein database and the common contaminants such as wool keratins and porcine trypsin, downloaded from Uniprot. The spectra were also searched against the Uniprot protein database, taxonomically restricted to chordates, and non-horse peptides were identified and eventually removed. Proteomic data were further compared to similar information already generated from fossil specimens collected in Siberian permafrost and temperate environments. Proteome-wide incidence of deamidation was estimated in relation with protein recovery to further assess the molecular state of preservation of ancient proteins.

Phylogenetic analyses. The CDS of protein-coding genes were selected from the Ensembl website, keeping the transcripts with the most exons in cases where multiple records were found for a single gene. We then extracted corresponding genomic coordinates, filtered for DNA damage/sequencing errors, and aligned each gene using MAFFT G-INS-i ('ginsi')^{48,49} (Supplementary Information, section 8.3a). Phylogenetic analysis was carried out using a super-matrix approach. First RAXML v7.3.2⁵⁰ was run to generate the parsimony starting trees. The final tree inference was performed using RAXML-Light v1.1.1⁵¹ and one GTRGAMMA model of nucleotide substitutions for each gene partition (codon positions 1 and 2, versus 3). Node support was estimated using 100 bootstrap pseudo-replicates. Bootstrap trees were dated using 'r8s', using the PL method and the Truncated Newton (TN) algorithm, with a smoothing value of 1,000 (ref. 52), or using the Langley–Fitch (LF) method (Supplementary Information, section 8.3.c). The date of the root node was constrained to 4.0–4.5 Myr, the date of CGG10022 was fixed to 43 kyr, and the date of the Thistle Creek specimen was constrained to 560–780 kyr BP. We also performed phylogenetic analyses of whole mitochondrial

genomes (Supplementary Information, section 8.1), Y chromosome (Supplementary Information, section 8.2) and a series of topological tests using approximately unbiased tests as implemented in the CONSEL makermt program⁵³ (Supplementary Information, section 8.3b).

Demographic reconstructions. Past population demographic changes were reconstructed from whole diploid genome information using the pairwise sequentially Markovian coalescent model (PSMC)²¹ and excluding sequence data originating from sex chromosomes and scaffolds (Supplementary Information, section 9). For low coverage genomes (<20×), we applied a correction based on an empirical uniform false-negative rate. Three different generation times of 5, 8 and 12 years were considered in agreement with the range of generation times reported in the literature^{23,54–56}. Mutation rates were estimated using quartet genome alignments where the donkey was used as out-group (Supplementary Information, section 10.1c). We also reconstructed past horse population demographic changes by means of Bayesian skyline plots using the software BEAST v1.7.2 (refs 57, 58) (Supplementary Information, section 9.2). Complete mitochondrial genomes were aligned and partitioned as described in Supplementary Information, section 8.1b, and a strict clock model was selected. We ran two independent MCMC chains of 50 million iterations each, sampling from the posterior every 5,000 iterations. We discarded the first 10% of each chain as burn-in, and after visual inspection in Tracer v1.5⁵⁹ to ensure that the replicate chains had converged on similar values, combined the remainder of the two runs.

Population split. We followed the method presented in ref. 20 to estimate the population divergence date of ancient and modern horses (Supplementary Information, section 10.1). This method was also applied to date the population divergence of Przewalski's horses and domestic horses (Supplementary Information, section 10.2), as both our phylogenetic analyses and admixture tests supported those as two independent populations (Supplementary Information, sections 8.3 and 12). In this method, we focus on heterozygous sites in one of the two populations and randomly sample one of the two possible alleles (ancestral or derived) in the individual belonging to the first population. The number of times a derived allele is sampled (*F* statistics) can be used to recover a full posterior distribution of the population divergence time using (serial) coalescent simulations and approximate Bayesian computation (ABC) (Supplementary Information, section 10.1). For dating the divergence time between the Przewalski's horse population and domestic breeds, we also performed coalescent simulations using ms⁶⁰ assuming different divergence times in order to compute the expected relative occurrences of 4 genotype configurations (Supplementary Information, section 10.2b). We assumed that no gene flow occurred after the population split, in agreement with the absence of detectable levels of admixture. The divergence time was then estimated by minimizing the root mean square deviation (r.m.s.d.) between observed and expected genotype configurations. We minimized the r.m.s.d. using a golden search algorithm. We repeated the minimization from different starting values to ensure convergence.

Selection scans. We used quartet alignments including the donkey as out-group, one ancient horse and two modern horses to scan for genomic regions where the two modern horses shared unusual accumulation of derived alleles (Supplementary Information, section 11.1). We used a sliding window approach on the entire genome, with a window size of 200 kb and calculated an unbiased proxy for selection using the 'delta technique' (see for example ref. 61). We then used an outlier approach to identify candidate loci with a conservative false-positive rate of 0.01. We further retrieved transcript IDs from the different genomic regions identified and performed functional clustering analyses in DAVID⁶². We estimated genetic diversity (theta Watterson) within the Przewalski's horse population and among modern horse breeds using sliding windows of 50 kb. For this, we estimated the population scaled mutation rate and used an empirical Bayes method where we took the uncertainty of the data into account by using genotype likelihoods instead of calling genotypes. We computed the genotype likelihoods assuming a model similar to that of SAMtools version 0.1.18 (ref. 37) (Supplementary Information, section 11.2). Genomic windows showing excessive proportions of segregating sites with regards to species divergence (>5%) or coverage <90% were discarded. We estimated Tajima's *D* following the same procedure and identified genomic regions showing minimal Tajima's *D* values and low genetic diversity among breeds but not in the Przewalski's horse population as a conservative set of gene candidates for positive selection among modern horse breeds. Finally, we scanned modern horse genomes for long homozygosity tracts, which could be indicative of selective sweeps⁶³. We used 2-Mb sliding windows and ignored sites showing coverage inferior to 8. This resulted in the identification of 456 outlier regions within 8 modern horse genomes.

Admixture analyses. In order to investigate if there was evidence for gene flow between the Przewalski's horse population and four modern horse domestic breeds (Arabian, Icelandic, Norwegian fjord and Standardbred), we performed ABBA-BABA tests^{20,24}. To avoid introducing bias due to differences in sequencing

depth we based the tests on data achieved by sampling one allele randomly from each horse at each site. First we used the domestic donkey as out-group, then the Middle Pleistocene Thistle Creek horse. When using the Thistle Creek horse as out-group we removed all sites showing transitions to avoid spurious patterns resulting from nucleotide misincorporations related to post-mortem DNA damage. We estimated the standard error of the test statistic using 'delete-m Jackknife for unequal m' with 10-Mb blocks⁶⁴ (Supplementary Information, section 12.1). We also scanned genome alignments to record the proportion of shared SNPs between Przewalski's horse and each horse breed (Supplementary Information, section 12.6), a proxy for recent admixture events that are expected to result in the introgression of alleles from the admixer to the admixed genome and long tracts of shared polymorphisms. Finally, we compared our Przewalski's horse individual to other individuals with different levels of admixture in their pedigree. We extracted genotype information from the Przewalski's horse genome for SNP coordinates already genotyped across 9 Przewalski horse individuals³⁸. Genotypic information from two Mongolian horses was added as out-group. We next selected the best model of nucleotide substitution using modelgenerator v0.85 (ref. 65) and performed maximum likelihood phylogenetic analyses using PhyML 3.0 (ref. 66) (Supplementary Information, section 12.5). We further confirmed the phylogenetic position of our Przewalski's horse individual together with Rosa (KB3838), Basil (KB7413) and Roland (KB3063), three individuals for which no admixture with domestic horses could be detected in previous studies²⁵ by means of Approximate-Unbiased (AU) and Shimodaira-Hasegawa (SH-) tests, as implemented in CONSEL⁵³.

Morphological analyses. We measured the metapodial of Thistle Creek early Middle Pleistocene bone for 6 dimensions, despite incomplete preservation of its distal end (Supplementary Information, section 1.2). These measurements were compared to 30 metatarsals of *E. lambei*, 9 metatarsals of *E. cf. scotti* of Klondike, Central Yukon, Canada (Supplementary Information, section 1.2) and to extant horses (Supplementary Information, section 1.3). Comparisons were made using Simpson's ratio diagrams that provide a standard and accurate comparison of both size and shape, for a single bone or a group of bones (Supplementary Figs 1.2 and 1.3). We also measured taxonomically informative morphometric features on the skull and post-cranial complete skeleton of the modern Przewalski's horse specimen that was genome sequenced. We compared those to a collection of horse measurements available for horses, filtering for specimens of similar age and using principal component analyses (Supplementary Information, section 1.4).

- Orlando, L. *et al.* Revising the recent evolutionary history of equids using ancient DNA. *Proc. Natl Acad. Sci. USA* **106**, 21754–21759 (2009).
- Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nature Protocols* **2**, 1756–1762 (2007).
- Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **6**, <http://dx.doi.org/10.1101/pdb.prot5448> (2010).
- Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).
- Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
- Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* **315**, 207–212 (2007).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- McCue, M. E. *et al.* A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.* **8**, e1002451 (2012).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
- R Development Core Team. A language and environment for statistical computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, 2011).
- McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
- Smith, C. L., Chamberlain, A. T., Riley, M. S., Stringer, C. & Collins, M. J. The thermal history of human fossils and the likelihood of successful DNA amplification. *J. Hum. Evol.* **45**, 203–217 (2003).
- Ginolhac, A., Rasmussen, M., Gilbert, T. M., Willerslev, E. & Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153–2155 (2011).
- Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104**, 14616–14621 (2007).
- Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26**, 1367–1372 (2008).
- Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

48. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
49. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
50. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
51. Stamatakis, A. *et al.* RAxML-Light: a tool for computing Terabyte phylogenies. *Bioinformatics* **28**, 2064–2066 (2012).
52. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
53. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
54. Lippold, S., Matzke, N. J., Reissmann, M. & Hofreiter, M. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol. Biol.* **11**, 328 (2011).
55. Achilli, A. *et al.* Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl Acad. Sci. USA* **109**, 2449–2454 (2012).
56. Warmuth, V. *et al.* Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc. Natl Acad. Sci. USA* **109**, 8202–8206 (2012).
57. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
58. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
59. Rambaut, A. & Drummond, A. J. Tracer v1.5. <http://beast.bio.ed.ac.uk/Tracer> (2009).
60. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
61. Zhang, Z. *Computational Molecular Evolution* (Oxford Univ. Press, 2006).
62. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* **4**, 44–57 (2009).
63. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
64. Busing, F. M. T. A., Meijer, E. & Van Der Leeden, R. Delete-*m* Jackknife for Unequal *m*. *Stat. Comput.* **9**, 3–8 (1999).
65. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).
66. Guindon, S. *et al.* New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

Chapter 8

Analysis of polar marine environments

This chapter addresses the analyses of the polar marine metagenomes. Figure 8.1 illustrates a general workflow of the involved steps. After assembling the metagenomes, genes were identified (described in Chapter 5 and 6 respectively) and a non-redundant gene catalogue was created. Proteolytic enzymes were identified from the nonredundant gene catalogue. Sequencing reads were remapped to the gene catalogue to generate the abundance matrix for further analysis.

8.1 From Genes to Abundance matrix

The non-redundant gene catalogue of the marine environment

A non-redundant gene catalogue can be created by pooling all identified genes from all samples and reducing sequence redundancies. Reducing redundancy in the polar marine gene catalogue was done with CD-Hit [101]. The basic principle of redundancy reduction lies within multiple sequence alignment and keeping the bin-representative gene (commonly the longest). The polar marine gene catalogue was binned with an identity cutoff of 95% over the shortest gene. Furthermore, cluster representatives shorter than 100 bp were removed. The number of genes was reduced from >13 million genes to ~5 million genes. The gene catalogue was used in the analysis of Manuscript II and III.

Abundance measure of metagenomes

Measuring the abundance of a given gene is important when comparing metagenomes. This can be done by aligning the sequencing reads to a non-

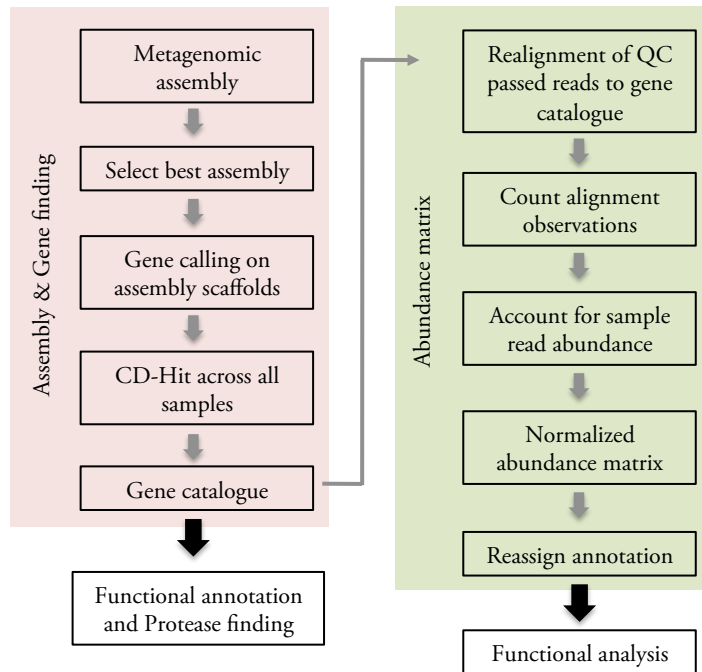


Figure 8.1. Workflow for the analysis of the polar marine environments

redundant gene catalogue of the metagenomes.

The quality trimmed reads were remapped to the non-redundant catalogue with the Burrows-Wheeler aligner mapping tool BWA [97]. BWA is a read alignment package which is based on the backward search with Burrows-Wheeler Transform.

Measuring abundance by remapping of reads singletons and paired-end reads have to be treated differently when assigning observations. An observation can be described as an observation count measure. If a singleton, a paired read or both paired reads mapped to a gene in the catalogue, the gene was observed once in the process. However, if a paired read was mapped to two different genes, both genes were observed once. Analyzing reads mapped to a reference is commonly done with samtools [98]. The sam format provides specific flags which can be specified with samtools to extract mapped reads from the bam formatted mapping file [98]. However, errors in BWA generated bam file flags were observed and therefore a customized perl script was used to read the alignment file line by line and determine the mapped gene and assign the gene observation count as described above.

The gene observations were done sample-wise, resulting in a $N \times M$ abundance matrix (N : number of genes in the gene catalogue, M : number of samples).

For further analysis the gene catalogue was normalized for the overall read count in the metagenomes, i.e. size of the different samples. This means that the samples are downsized to the most low abundant sample (Equation 8.1).

$$a_{downsized} = a * A / A_{min}$$

$a_{downsized}$: downsized gene abundance

a : gene abundance

A : sample read count

A_{min} : min sample read count

(8.1)

Downsizing, however, does impact the overall resolution of the abundance of samples which means that deeply sequenced samples will be downsized to a scale where small differences in abundance cannot be detect. This becomes most apparent with samples of varying sequencing depth. Unfortunately this was the case with the polar marine metagenomes. Excluding samples with lower sequencing depth would have reduced the sample size dramatically since only 26 samples were available. In studies with more samples [6, 131], however, this might very likely be a good option. For the functional analysis of the polar marine environment, only 25 samples were used as one sequencing lane was sequenced with a mixture of two individual samples and post separation of the samples was not possible. However, for the identification of proteolytic enzymes, all 26 samples were used.

The same principle of remapping reads to the non-redundant gene catalogue was used for creating rarefaction curves of the samples. Reads were randomly chosen and mapped to the catalogue with BWA [97]. Optimally the rarefaction curve reaches a plateau indicating that by adding more sequence information, no more genes from the gene catalogue can be detected in the specific sample.

8.2 Functional analysis – Finding the needle in the haystack

The 25 samples, which were used in the analysis of Manuscript II, were divided into groups, which represent surface (40 m – 100 m), medium (300 m – 400 m) and deep (2,000 m – 4,300 m) environments of true open ocean, where the deep samples were taken relative to the ocean bottom representing the whole water column. Moreover, the archipelago samples represent a coastal environment with varying depths (400 m – 1,500 m). In order to get a measure of similarity between the four environments, the Bray–Curtis dissimilarity¹ measure was calculated between the samples using the abundance

¹The non-metric Bray-Curtis dissimilarity [19] delivers robust and reliable dissimilarity results for a wide range of applications. It is a commonly applied measurement to express relationships in ecology and environmental sciences [17].

matrix.

Finding meaningful functions in metagenomic data is not a trivial task due to the vast amount of data. The gene abundance matrix was translated to COG and NOG functions according to the genes annotation. The resulting abundance matrix had the dimension of ~5 million genes x >13 thousand functions. The mean abundance for each COG/NOG function was plotted. The Kruskal–Wallis test¹ was applied on a subset of the new abundance matrix to statistically compare the surface, medium and deep samples. Sample subsets were compared by calculating the negative mean log-ratio for samples of the same category. A cutoff was set to identify functions which were more prominent in a sample category.

Statistical analysis narrowed down the number of functions to manually look at. However, the identification of interesting functions, still required manual curation.

8.3 Identifying proteolytic enzymes in the polar marine environment

The non-redundant gene catalogue was used to identify proteolytic enzymes in Manuscript III. All ~5 million genes were annotated with HMM models created in collaboration with Novozymes. In total, 3,207 HMMs were constructed and used in the screen of the non-redundant gene catalogue.

These HMM models include all peptidases defined in MEROPS [136] and were constructed based on individual peptidases in MEROPS (e.g. A02.063), using nearby homologues in known sequence space (UniProt [8] and Novozymes' internal protein database). Up to 500 nearby homologues were identified using BLASTP [4] against the peptidase's catalytic domain. A given protein could only be assigned to the closest peptidase in MEROPS, and thus never used in multiple models. The catalytic domain of each peptidase was extracted and a multiple alignment was created using MAFFT [83]. The HMMs constructed using HMMER3's hmmbuild [41].

The identified proteases were further aligned to public databases (UniProt [8]) and metagenomic datasets to identify novel sequences and narrow down the number of novel targets for future expression trials. The metagenomes were downloaded from CAMERA [150]. Moreover, signal peptides were identified with SignalP [125].

¹The Kruskal–Wallis test [89] compares the medians of two or more samples to determine if the samples originate from different populations. Adapted from Spurrier [156].

Part III

Manuscripts

Chapter 9

Manuscript II

9.1 The polar marine environment

Oceans cover ~70% of the Earth's surface and contains 97% of the planet's water. They play a pivotal role in many of the Earth's systems including climate and weather¹. The marine environment is considered to be on of the largest habitats on Earth with 2.9×10^{27} cells in deep water (below 200 m) to 3.6×10^{28} cells in surface water (above 200 m) accounting for 55% of all prokaryotes in aquatic habitats [179]. These numbers suggests that oceans account for one of the largest biomes on Earth. However, more than 95% of the underwater world remains unexplored [129, 179]. Interest in the marine environment has long been a subject of fascination due to its alien character compared to terrestrial environments. Especially the ever dark deep sea has been an interesting topic for research. The first scientific evidence for life in the deep seas was found in the late 1800s by the Challenger Expedition [54, 75, 79]. Since then numerous expeditions have been exploring the world's seas to uncover the biological diversity, and microorganisms have been identified in niche environments such as hypothermal vents, deep sea, on whale carcasses and Arctic waters [79, 154, 179]. One of the most extensive marine sampling expedition was conducted in 2007 – 2009 known as the Global Ocean Sampling Expedition [122, 140] where up to 400 liters of water were sampled approximately every 200 miles. The metagenomes of the samples are publicly available providing a "global map" of the microbial diversity throughout the globe.

Marine bacteria (and also archaea) are an integral and important part in the biogeochemical cycles by steadily assimilating, storing, transforming, exporting and remineralizing the vast pool of organic carbon stored in oceans

¹<http://www.noaa.gov/ocean.html>, accessed 18. November, 2013.

[179]. The microbial composition is highly diverse and creates a very complex ecosystem, driving the main biological processes in the marine ecosystem and conducting various and diverse metabolic functions like photosynthesis, CO₂ fixation, heterotrophic processes and utilization of inorganic compounds [163, 186, 193].

The Arctic and Southern oceans are considered extreme environments due to low surface temperatures, general low nutrition and ice coverage. The global thermohaline circulation connects both oceans and it takes approximately 1,000 years for water masses to circle the globe through the deep current system [183]. A recent study [159] showed, however, that the polar microbial communities in the Arctic oceans and Southern oceans follow a bipolar distribution, i.e. similar species distribution at polar regions separated by the moderate zones.

In the Arctic oceans little temperature changes have been observed [24, 78]. However, various other environmental factors differentiate surface and deep water throughout the water column. One of the most prominent environmental changes is the sunlight penetration. The photic zone is roughly 200 meters deep and provides energy to organisms metabolizing inorganic compounds (primary producers) and photoheterotrophic organisms. Microbial growth is influenced by sunlight and it has a selective influence on the community structure [29]. Another differentiating factor along the water column is the organic matter influx described as dissolved organic matter (DOM) [86]. DOM is categorized in three groups varying along the depth gradient [86]. Labile DOM consist of proteins, free amino acids and sugars and exhibit a low concentration of $< 1 \mu\text{mol/L}$. The semi-labile DOM's concentration decreases from max. $30 \mu\text{mol/L}$ at the surface to zero at a depth of approximately 1,000 m [119]. The typical degradation time ranges from days to years and the processes involved are still unknown [119]. The refractory pool of DOM has a degradation time of over 1,000 years exceeding the deep water circulation and creates a constant concentration of $40 \mu\text{mol/L}$. Particulate organic matter (POM) also play a prominent role in the ocean nutrition cycle because it enables nutrients to sink to deeper levels of the water column [69, 86]. Previous studies have shown that the depth gradient is important in shaping the microbial communities in the oceans [37, 57].

Here we present the functional analysis of water specimens sampled during the Galathea III and LOMROG II polar expeditions (Figure 9.1). DNA was extracted from 26 samples, sequenced and assembled. A non-redundant gene catalogue was created. All genes of the polar marine catalogue were assigned to their function. The functional stratification according to depth and north/south was investigated for 25 samples – one sample was excluded from the functional analysis as one sequencing lane was loaded with a mixed sample. We find that functional stratification occurs along the water column but to a lesser extend in the geographical orientation. The current version of the manuscript is appended.

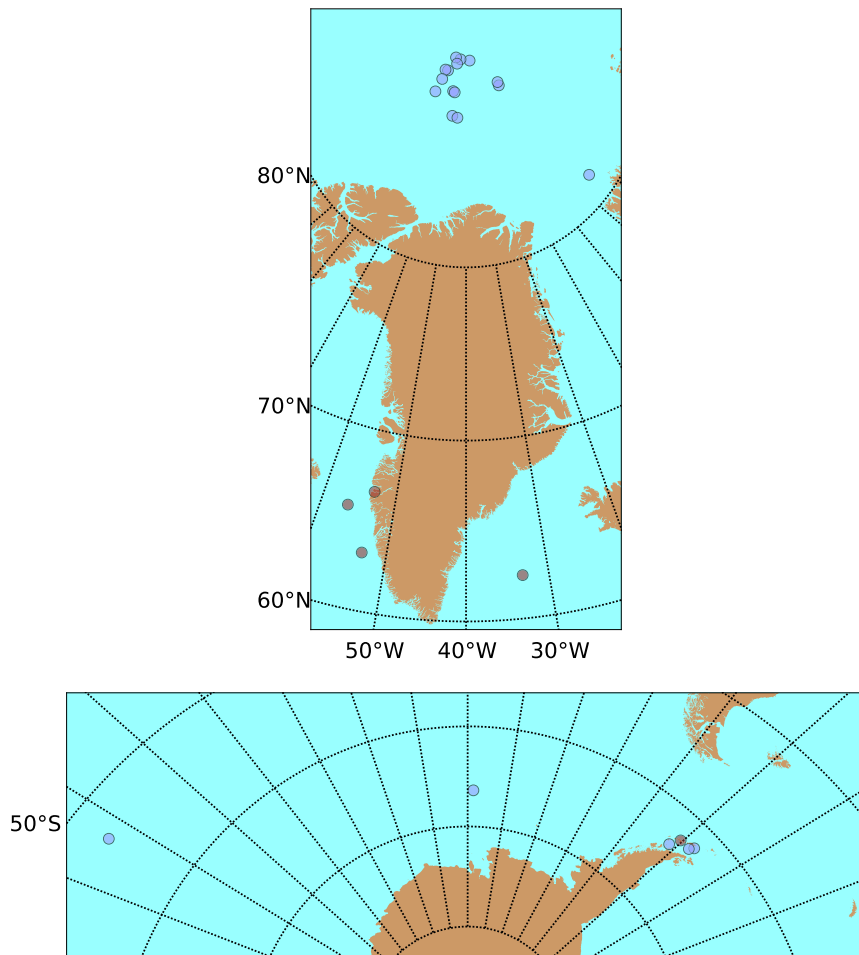


Figure 9.1. Map of sampling site of the Galathea III and LOM-ROG II polar expeditions in the Arctic (top) and antarctic (bottom). Blue dots represent sample locations, red dots in the top represent test sampling sites which were not sequenced. Illustrations were modified from Hansen et al., unpublished.

9.2 Comparative functional analysis of Arctic marine metagenomes reveals strategies for deep sea persistence

Comparative functional analysis of arctic marine metagenomes reveals strategies for deep sea persistence

Josef Korbinian Vogt^{1*}, Lea Benedicte Skov Hansen^{2*},
Dhany Saptura¹, Peter Nikolai Holmsgaard², Lars Hestbjerg Hansen²,
Søren Sørensen², Thomas Sicheritz-Pontén¹ and Nikolaj Blom¹

¹Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark

²Department of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark

* Joint first authorship.

ABSTRACT – The global Ocean represents the worlds largest continuous ecosystem, however, little is known about the microbial functions present in the aphotic zone. The microorganisms that dwell in the deep dark waters are numerous and play key roles in the ocean carbon cycle, which is of special interest in connection to the global climate change. The polar oceans are highly affected by the temperature increase and this emphasizes the need for a better understanding of the biological processes present throughout the water column. The purpose of this study was to conduct a functional clustering of metagenome shotgun DNA libraries of samples from the Arctic Ocean and the Southern Ocean and through the water column reaching levels lower than 4,000 m. This dataset represents the deepest microbial samples from these oceans to date. Furthermore, a comparative analysis was conducted to infer the functional differences between the environments. The results indicated that the environmental factors differentiating through the first 300 m of the water column are deciding factors for shaping the functional community, rather than spatial dispersal. The mesopelagic samples were functional inseparable from the bathy- and abyssopelagic samples, indicating a highly homogenous environment in the aphotic part of the ocean. Functions characterizing the aphotic zone were iron uptake and utilization, phage and bacteria interactions, adhesion and motility and others, which in general indicated a selection for copiotrophs in the deeper ocean.

KEY WORDS – Metagenomics, polar marine environment, Arctic, Antarctic, deep-sea, functional analysis

1. Introduction

Approximately 70% of the Earth's surface is covered by ocean and contains 97% of the planet's water. The marine environment is considered to be one of the largest biomes on Earth with 2.9×10^{27} cells in deep water (>200 m) to 3.6×10^{28} cells in surface water (<200 m) [58]. However, more than 95% of the underwater world remains unexplored [48, 58], even though these microorganisms play a pivotal role in the world's carbon cycle. In connection to the global climate change, especially the polar oceans are undergoing environmental changes and this emphasizes the need for a better understanding of the biological processes conducted throughout the water column. The Arctic Ocean and the Southern Oceans can be described as extreme environments with low nutrition, low surface temperatures, ice coverage and 24 hour solar irradiation at summertime. Despite the similarities in environmental factors, both oceans exhibit geographical differences. The Southern Ocean encloses a continent and

the Antarctic Circumpolar Current (ACC) isolates its water masses. The lack of freshwater inflow keeps the level of salinity constant throughout the water column [44]. The Lomonosov Ridge divides the Arctic Ocean into the Mesozoic Amerasian Basin and the Cenozoic Eurasia Basin. The ridge creates a natural barrier, which prevents the basins' water to be mixed [10]. Moreover, the Arctic Ocean water is surrounded by Canadian, Russian and Greenlandic land masses and receives 10% of the world's fresh water. This causes reduced salinity in the surface water layer. Despite the geographical isolation of the Arctic and Southern Oceans, they are connected by the global thermohaline circulation allowing for water circulation between the poles. This, however, takes up to 1,000 years [61].

The temperature of the arctic oceans varies little throughout the water column. However, other environmental elements change from surface to deep waters, such as light penetration and organic matter concentration. Primary producers are mostly present in the upper 200 m where

light penetrates the water and their presence promotes a general higher cell density compared to deeper environments [58]. Availability of sunlight directly influences microbial growth and has a selective impact on the community structure [9].

Organic matter in oceans are present at varying concentrations and compositions along the depth gradient. Labile dissolved organic matter (labile DOM), including free amino acids, sugars and proteins, is only present through the photic zone in low concentrations of $<1 \mu\text{mol/L}$. Semilabile DOM exhibit a decreasing concentration with maximum $<30 \mu\text{mol/L}$ at the surface to zero at a depth of 1,000 m [42]. They can persist over 1,000 years, which exceeds the deep water circulation and creates a stable concentration of $40 \mu\text{mol/L}$. Particulate organic matter (POM) also play a prominent role in the ocean nutrition cycle because it enables nutrition to sink to deeper levels of the water column [40].

Microbial community studies have shown a vertical zonation of microbial communities in oceans [12, 26, 62]. Taxonomic comparison of surface and deep water from the Arctic Ocean and Southern Ocean did reveal a bipolar distribution of bacteria (Hansen *et al.*, *unpublished*) [18] by comparing the surface and deep samples of the Arctic and Southern oceans. Furthermore, previous studies of temperate oceans also suggest depth to be the determining factor for functional stratification [12, 26, 36, 56].

We hypothesize depth to be the determining factor for functional clustering of metagenomic samples of the Arctic Oceans and Southern Oceans. Furthermore, comparative functional analysis of metagenomes of the polar oceans reveals environmental strategies for the adaptation of the microbial community to the extreme environment. Water samples from the Arctic Ocean and the Southern Ocean were collected during the Galathea III and LOMROG II polar expeditions at varying depth. The samples represent the whole water column ranging from 40 m to 4,300 m. Moreover, water was sampled at the archipelago in close proximity to the Antarctic Peninsula. These samples represent a coastal environment compared to the open ocean samples. With these samples it is possible to statistically investigate the functional stratification according to depth and Arctic Ocean versus Southern Ocean.

2. Results

To conduct a comparative metagenomic analysis of the marine environment at the arctic poles and throughout the water column, 25 arctic marine water samples were obtained (see Figure 1 and Table S1). In connection with the LOMROG II expedition in August, 2009, eleven different locations in the Arctic Ocean were sampled, which yielded sixteen samples. Both sides of the Lomonosov ridge were sampled to represent the Mesozoic Amerasian

Basin and the Cenozoic Eurasia Basin [10]. Furthermore, all sample locations were maximum 350 kilometers apart and close to the North Pole. However the P20 sample was located near Svalbard and was more distant.

Five different locations were sampled in Southern Oceans during the Galathea III expedition in January, 2007, which yielded nine different samples. Two sample locations represented a true open ocean environment, where P10 was situated north from the ACC and P11 was positioned south from the current [44]. P12, P14 and P15 were all sampled within the archipelago near the Antarctic Peninsula.

The 25 samples represent different ocean environments, which are surface (40 m – 100 m), medium (300 m – 400 m) and deep (2,000 m – 4,300 m) of true ocean, where the deep samples were taken relative to the ocean bottom. Hence, the samples represent the whole water column. Furthermore, the archipelago samples represent a coastal environment with varying depths (400 m – 1,500 m).

The measured temperature and salinity correlate with the seasonally obtained data published in the World Ocean Atlas [3, 53]. The temperatures ranged between -2°C and 2°C , except the surface and medium samples of P10, located north from the ACC, which were around 7°C and 9°C respectively. The salinity of the arctic surface samples was highly affected by ice melting and fresh water inlet and displayed concentrations between 31 PSU and 33.5 PSU. All other samples displayed stable concentrations between 34 PSU and 35 PSU.

Metagenomic shotgun sequencing libraries of bacteria and archaea were obtained from the water samples by excluding larger eukaryotes and viruses with filtering. Libraries of paired-end sequences were obtained and the sizes ranged from 1.38 Gb to 25 Gb after quality trimming. However, the deep sample from P11 has been deeply sequenced which gave 37.71 Gb (Table S1). These libraries were assembled to contigs, which displayed an *N50* between 774 and 40,444. The number of contigs ranged from around 1,000 up to 700,000. The sequencing statistics imply a connection between lower sequencing coverage, higher *N50* and lower contigs and vice versa.

The metagenome shotgun sequences were used to assess the taxonomic distribution of bacteria and archaea in the samples. The results are displayed as pie charts in Figure 1, where the ten most abundant phyla in the individual samples are included. Proteobacteria were the most abundant phyla across the samples but also Actinobacteria, Firmicutes and Bacteroidetes were highly abundant and these phyla did not display any preferences for a specific depth or location except Actinobacteria, which showed a slight preference for deeper waters in the south. Thaumarchaeota and Chloroflexi showed a more clear tendency toward medium and deep water and Acidobacteria and Deinococcus-Thermus showed a preference toward deep and archipelago environments.

The data showed that a fairly large proportion of the assigned reads fall into the “other” category and this implied a long tail of rare phyla in the samples. Furthermore, between 19.56% and 85.01% of all sequences in the sample libraries could not be assigned to any phyla. This indicated that the samples were containing novel marine organisms that might also introduce a certain bias when assessing the taxonomic distribution.

Over 5 million genes were identified after assembling all 25 metageomes, gene calling and clustering. All reads were remapped to the non-redundant gene catalogue to create the abundance matrix. Furthermore, the reads were rarified and remapped. Figure 2 displays a functional rarefaction analysis, where number of genes from the gene catalogue are displayed as a function of number of reads in the individual samples. Samples from all four sample types, surface, archipelago, medium and deep, almost reached a plateau at around 2 million genes, hence they showed similar functional diversity. All deep samples displayed a similar curvature, however, surface, archipelago and medium samples were deviant within sample groups. This divergence could be explained by lower functional diversity, however some samples also had a smaller library size and contributed less to the gene catalogue, resulting in less functional coverage of the sample. This was particularly evident for surface and medium samples from the south.

Approximately 3 million genes could be annotated to over 16,000 ortholog groups (OGs) [39] and the gene abundance matrix was rearranged to create a functional abundance matrix. The Bray-Curtis dissimilarity measure was calculated between the samples and the result is displayed in an NMDS ordination plot, see Figure 3A [7]. All samples clustered within close proximity to each other and displayed a low functional diversity between the samples. Surface samples showed larger diversity and were distinct from the tight cluster of medium and deep samples. Within the tight cluster, the medium and deep samples from the North Pole were functionally inseparable, the two southern deep samples diverged slightly from northern cluster and medium samples from south were clustering together with the surface samples. The archipelago samples clustered together with surface samples and these two environments did not separate from each other on a functional basis. Metadata, including depth, temperature, salinity and geographic location, was fitted as vectors onto the ordination plot. Depth and location affected the sample clustering in the NMDS plot with statistical significance (p-values of 0.00281 and 0.00195 respectively). Salinity also showed statistical significance (p-value of 0.02711), however to a lesser extent. Temperature seemed not to affect the ordination and was not statistically significant.

In order to compare the functional composition throughout the water column all OGs found in surface, medium and deep samples were displayed in a ternary plot (see

Figure 3B). The distribution of the functions between the three environments created two broad clusters. One cluster moved vertically up through the middle of the plot toward the surface environment and a second cluster spread out across the bottom of the triangle. The vertical cluster in the plot represented many functions, which were shared between all three sample types. The cluster were slightly skewed towards the medium environment, hence surface samples displayed an environment more similar to the medium environment than the deep. However, only few OGs were uniquely shared between surface and medium. The upper corner of the triangle was populated by many data points and the associated functions displayed a unique preference for the surface environment.

The bottom of the triangle represented a low abundance in surface and the cluster consisted of functions present mostly in deep and medium samples, with a slight trend towards medium samples. Almost all functions in this cluster were equally shared between medium and deep and only few OGs favored a specific depth, which was evident by the few data points present in the lower corners of the triangle.

To further investigate the functional diversification throughout the water column, OGs over 90% and less than 5% present in surface samples compared to medium and deep samples were extracted from the ternary plot and a statistical Kruskal-Wallis one way analysis of variance was utilized to identify statistic significantly different abundant OGs between surface, medium and deep [28]. Approximately 47% of the identified OGs, had no known function associated and were excluded from the dataset. In total, 176 diverse functions were identified and these are displayed in a heatmap, see Figure S1, where 26 OGs exhibited higher abundance in surface samples and 150 OGs displayed higher abundance in the medium and deep layers. Clusters of OGs that preferred either the medium or deep environment were not detected.

The OGs were divided into functional subgroups and three of the most prominent were iron uptake and utilization, phage and bacteria relations and adhesion and motility, which are displayed as heatmaps in Figure 4. Iron uptake and utilization seemed to be higher abundant in deeper samples (see Figure 4A). OGs like isochorismatase hydrolase (NOG138795) and TonB related proteins including the ExbD/TolR biopolymer transport protein (NOG121145, NOG243700 and NOG252395) were identified. The isochorismatase hydrolase conducts an important step in the siderophore synthesis [19]. These siderophores binds extracellular ferric iron and are then transported into the cell via the TonB transport system to provide the organism with vital ferric iron from depleted surroundings [41, 60]. Three different OGs with heme binding proteins were also identified (NOG74099, NOG118022 and NOG83915). Extracellular heme binding proteins bind heme molecules outside the cell and then enters the cytoplasm via an ABC transport system [37]. Furthermore, heme containing pro-

teins binds O₂, NO, H₂S and CO and are key players in respiration and signalling [37]. The identified protoporphyrinogen oxidase (NOG145956) and uroporphyrinogen-III decarboxylase (NOG72702) participate in the bacterial heme biosynthesis, which supplies the heme molecules for metalloproteins. Of other deeper metalloproteins, a hemerythrin HHE cation binding protein (NOG145840), Quinohemoprotein amine dehydrogenase (NOG73521) and a NADH-quinone oxidoreductase (NOG145840) were found. Hemerythrin has been shown to respond to NO and oxidative stress [8, 45, 50]. Furthermore, the quinohemoprotein amine dehydrogenase participates in degrading complex amines in the periplasm to provide electrons for energy production [17]. In general, iron-sulfur clusters are important participants in various metabolisms in cell. The identified NADH-quinone oxidoreductase contains eight or nine clusters and participates in electron transport chains. Also, a protein involved in iron sulfur cluster assembly (NOG242690) have been identified. Figure 4B displays OGs which have connections to phage and bacteria relations. The phage-related lysozyme (COG 3772) was the only phage related function found significantly for the surface environment. These types of proteins are participating in the lytic cycle, where the phage lyses the host cell to release the newly synthesized phages [29]. In the deeper waters, functions pointing towards a lysogenic phage strategy were present. The Mu-like prophage FluMu protein gp28 (COG4373) indicated the presence of a Mu-type prophage, who integrates into bacterial genomes. This integration occurs through proteins involved in DNA integration (NOG10655) and transposases (COG5659, NOG247136, NOG149091 and NOG 236938), however transposases participate in general lateral gene transfer of transposons and are not only connection to prophages. OGs that participate in bacteriophage defence systems were also identified in the medium and deep samples. To prevent phages from recognizing specific receptors on the bacterial cell wall, exopolysaccharides block the phage recognition sites, where capsular polysaccharide synthesis proteins (NOG41724) are responsible for producing these [29]. Furthermore, CRISPR-Cas systems associated functions (COG3513 and COG3649) were present in the deeper samples, which are identified as phage defence mechanisms [29]. Abortive infection proteins, represented by two OGs (NOG138780 and NOG10149), participate in a different defence mechanisms, where the proteins target crucial steps in the phage multiplication, which often leads to cell death [29]. One of the well known defence systems against bacteriophages are the restriction modification systems, where the host DNA is methylated by methylase/methyltransferase and recognized by endonucleases and restriction enzymes, which will degrade foreign non-methylated DNA [29]. Several methylases/methyltransferases, restriction enzymes and endonucleases (COG0827, COG1002, NOG83182, NOG246631, NOG45993, NOG81569, COG4797 and

NOG09292) were identified in the medium and deep data. However, methylases/methyltransferases and endonucleases have various biological functions and could participate in other pathways than the restriction modification system.

OGs with connections to motility and adhesion also seem to be prominent in the deeper waters (see Figure 4C). Alpha integrin proteins (NOG146018) are well characterized adhesion mediators, which contain the extracellular FG-GAP repeat region acting as specific adhesion recognition sites [47, 52]. Proteins involved in biological adhesion and cell-cell adhesion (NOG149619 and NOG245115) also provide receptors directed towards specific adhesion. These receptors are often carbohydrate binding proteins (NOG85828), also called lectins, where the specific carbohydrate are the actual recognition site [2]. Furthermore, two potential membrane bound lipoproteins were identified (NOG76757 and COG1724). They can act as adhesins, however they possess many diverse biological functions [59]. The lipopolysaccharide kinase (NOG42907) and the proteins involved in positive regulation of lipoprotein lipase (NOG81106) are key enzymes in lipoprotein synthesis [38, 63]. Capsular polysaccharide proteins (NOG41724), mentioned earlier to be important in phage defence systems, also play a role in adhesion, where the extracellular polysaccharides act as receptors and as a sticky mass that will adhere to various surfaces [2]. Furthermore, adhesion molecules with a Ig like domain (NOG 257069) have been shown to mediate intimin cell-cell adhesion [23]. In connection with motility, an OG for an anti-activator of flagellar biosynthesis, FleN (COG0455), has been identified in the deeper samples. This function has been shown to regulate polar flagella synthesis [11]. Besides functions connected to iron uptake and utilization, phage and bacteria relations and adhesion and motility, OGs associated to carbon source and energy metabolisms were identified (Figure S1). In the surface, OGs describing metabolisms of small sugar molecules like trehalose, maltose and lactose (COG1554 and NOG130892) were identified, whereas functions participating in degradation of larger polysaccharides like polygalacturonan, hemicellulose, cellulose, starch, pectin and alginate were found in medium and deep samples (COG5434, NOG 70431, NOG147608, NOG71025, COG4692, NOG04112, COG3387, NOG45527 and NOG39328). Evidence of methanogens was also found in the deeper samples. The formylmethanofuran dehydrogenase (NOG1153) is a molybdenum binding iron-sulfur protein that participates in the reduction of autotrophic carbon dioxide [5]. Autotrophic lifestyle and energy production by sulfur oxidation was indicated by the oxidation protein, SoxZ (NOG19503) [16]. At last, the deeper samples contained many functions connected to DNA repair (COG2094, NOG39498, NOG 135388, NOG09685, COG1669, COG3298 and NOG0929) and regulation and signaling (NOG71309, NOG237274, NOG252009, COG0631, NOG235513, NOG252472,

NOG85867, NOG148303, NOG242754, NOG14654, COG2319, NOG68923, COG4455, NOG149909, NOG146568 and NOG71620), where functions participating in photosynthesis were identified in surface samples (NOG04867, NOG04871, NOG05023, NOG06447 and NOG08121) (see Figure S1). To infer the differences between the coastal and open ocean environment, the three archipelago samples were compared to the six true open ocean surface, medium and deep samples from the south. The ratio of the mean abundance of the OGs are displayed in a ratio plot (see Figure 5A). The 26 OGs with log ratios >4 were classified as being higher abundant in true ocean and the 1350 OGs with log ratios <-4 were classified as being higher abundant in coastal waters. Hence, many functions were characterizing the coastal environment and a large proportion of these had zero abundance in the true ocean samples. An interesting group of functions found in the coastal environment were polysaccharide degrading enzymes that act on xylan, alginate, agarose, hemicellulose, glycogen and pectin (NOG14217, NOG139342, NOG10914, NOG84929, NOG05353, COG3408, NOG73253 and NOG82826). Furthermore, functions involved in denitrification and sulfur oxidation were found including small indications of aerobic nitrogen fixation (COG4263, NOG133395, NOG123068, NOG78981, NOG72070, NOG13638, NOG118539 and NOG19503). Otherwise, many of the coastal OGs assigned to unknown or general functions like cytochromes, ribosomal proteins and general transporters. No functional pattern could be inferred of the 26 true ocean OGs. The two deep samples from the south pole (P10D and P11D) were deeply sequenced and suitable for making a comparison between the two poles. P22D and P23D were chosen for the comparison, where the average abundance was compared between northern and southern samples and displayed in a ratio plot, as described for the comparison between open ocean and coastal samples (see Figure 5B). Based on the distribution of the data, OGs with a log ratio >2 were classified as abundant in deep South Pole samples and ratios <-2 were classified as abundant in deep North Pole samples. The ratio plot created a more symmetric pattern and the number of GOs assigned to a specific geographic location were fairly similar with 928 OGs abundant in northern samples and 535 OGs abundant in the southern samples, where the total number of OGs is 8,068. Most of the functions assigned to the identified OGs were of general character or unknown and southern functions complemented functions found in northern samples. Hence, it was not possible to identify potential functional subgroups that differentiated the two geographic locations. However, the distribution of Mu-phage functions indicated the presence of specific phages at the North Pole (COG4228, COG4373, COG4396, COG4388, COG4397, COG3941, COG4382, COG4383, COG4387, COG5003, COG4384, COG4381, COG5005, COG4379, COG4386).

3. Discussion

To our knowledge, this work represents the first study of the whole water column of the Arctic Ocean and also includes reference samples from the Southern Ocean using a metagenomic shotgun sequencing approach. A comparative analysis of functions that persisted in epipelagic, mesopelagic and bathy- to abyssopelagic zones of the Southern Ocean and Arctic Ocean revealed a diverse surface layer compared to a functional inseparable medium and deep environment at the north pole (Figure 3). Surface and deep samples from the two poles clustered together and indicated that the environmental factors through the water column decides the functional composition of these marine micro communities despite the large geographic separation. An enquiry of depth specific OGs revealed functional traits like photosynthesis in the surface but also information on iron uptake and utilization, phage and bacteria relations and adhesion and motility (Figure 4). The taxonomic distribution of microorganisms in the samples was investigated to reveal the composition and potential stratification through the water column and from pole to pole. All phyla identified in the taxonomic analysis are well known for their presence in ocean habitats and the data did not reveal any pole specific phyla [6, 12, 15, 36, 49]. The abundance of Actinobacteria, Thaumarchaeota, Chloroflexi, Acidobacteria in medium, deep and archipelago samples indicated a taxonomic stratification through the water column. These result correlated with the findings from the 16S rDNA gene amplicon analysis conducted on the same DNA samples (Hansen *et al.*, unpublished). However, the 16S rDNA gene amplicon analysis revealed a higher resolution of a depth specific community structure in the Arctic Ocean, which is also confirmed by other studies [4, 18, 20]. The assessment of the taxonomic distribution from metagenomic data is highly dependent on the individual genome sizes as smaller genomes will contribute less to the abundance measured. Furthermore, no existing database covers all microbial genomes in the ocean and this is evident from the many unknown reads in the data (unknown reads made up 19.56% to 85.01% of the samples) and probably, the database is biased toward well researched phyla. Hence, these parameters might introduce bias in the abundance assessment. However, the metagenomic data identified phyla which were not reported by Hansen *et al.*, unpublished, like Firmicutes, Spirochaetes, Fusobacteria, Denioccocus-Thermus and Tenericutes, which indicate that the metagenomic classification of taxa circumvent the primer specificity bias introduced in a 16S rDNA gene amplicon analysis, however the abundance measure suffers from the sparse coverage of whole genome sequences of marine organisms. The functional rarefaction analysis did not indicate any differences in functional richness between surface, archipelago, medium and deep samples. Even though the cell density in deeper water is ten times lower than at the sur-

face and nutrition is very sparse [42, 58], the organisms residing in the meso- to abyssopelagic zones display an equally complex metabolic composition within the sample as the surface community. All samples seem to saturate around 2.5 million genes per sample and this is evident by looking at the deeply sequenced P11D, which displays the same curvature as the other deep samples, however, with twice the sequence depth. Some archipelago, surface and medium samples saturates at lower gene counts and this is tightly connected to the genes called in the assembly and a higher curvature might be observed with deeper sequencing. This means that the resolution for these samples might be too low to observe all depth specific functions in the comparative analysis.

The NMDS plot indicated patterns of functional stratification through the water column (Figure 3A). The surface samples created a diverse cluster compared to the tight medium and deep cluster and the southern medium samples appeared in an intermediate zone between surface and deeper water samples. This suggests that the functional differentiation through the water column in the Arctic Ocean is most prominent the first 300 meters and only limited differentiation occurs between the mesopelagic and bathy- to abyssopelagic zone, regarding the survival and persistence strategies for marine microorganisms. This is also evident when observing the ternary plot, where only few functions show a unique preference for either the medium or deep samples (Figure 3B). Furthermore, the Southern Ocean surface and deep samples clusters with samples of the same depths from the Arctic Ocean and this suggest that environmental factors are the directing parameter for functional composition of a microbial community rather than dispersal distance.

Functional depth stratification is also found in more temperate oceans citeDeLong2006,Konstantinidis2009b,Martin-Cuadrado2007b,Thureborn2013, however, a differentiation between the bathypelagic and mesopelagic layer was also observed [12]. Martin-Cuadrado *et al.* [36] finds bathypelagic samples from the warmer mediterranean sea to resemble subtropical Pacific Ocean mesopelagic samples and suggests temperature as the deciding factor for this functional clustering [12]. Since the temperature differences in the Arctic Ocean are limited, this might explain the clustering of the medium and deep samples.

The 16S rDNA gene amplification analysis by Hansen *et al.*, unpublished showed a clear phylogenetic stratification through the water column. Combined with the current functional results, this indicates that deep or medium specific organisms display similar means of survival in the meso- to abyssopelagic zones. In general, the functional metagenomic data display a tighter clustering compared to the phylogenetic clustering (Hansen *et al.*, unpublished). This indicates a core set of functions that are shared between the three environments and these are evident from the ternary plot, where many common functions can be observed. These common functions are well rep-

resented in the gene catalogue and with more sequencing depth, lower abundant functions might be revealed and a better differentiation between the samples might be observed.

The depth specific OGs were selected in a two step process, where differential abundance was ensured by calculating the mean abundance ratio between surface and deeper samples. Furthermore, a uniform presence over the samples and a statistical significant difference was calculated by the Kruskal-Wallis test, to exclude all OGs only present in few samples and not representative for the epipelagic or meso- to abyssopelagic environment. More OGs, which represent the medium and deep environment, were identified compared to unique surface OGs. The medium and deep samples represents a more homogenous environment and this possible yielded more common OGs. Furthermore, many of the OGs identified in the metagenomes were of unknown functions, hence, the subgroups found in this data are not representing the whole picture of a deep sea lifestyle and many more mystery functions need to be resolved.

Not surprisingly, OGs of photosynthesis functions were characterizing the surface community. Otherwise, iron uptake and utilization seemed to be important characteristics for deeper environment. The dissolved iron profile of the Arctic Ocean separates from other oceans because it is highly affected by the freshwater inflow and the presence of continental shelves, which supplies the surface water with high iron concentrations [24]. However, the pronounced halocline prevents the ocean water from mixing and this creates a iron depleted environment in the deep sea, while the high amounts of iron resides in the upper layers [24, 25]. This dissolved iron stratification might explain the abundance of iron uptake apparatus in deeper samples.

Carbon is another growth limiting factor in the ocean and there is some evidence of autotrophic lifestyles, especially in the deeper samples. However, enzymatic functions involved in polysaccharide degradation were abundant in the deeper samples, whereas evidence of smaller sugar metabolism was identified in the surface. The presence of polysaccharide degrading enzymes in deep marine organisms indicates a copiotrophic lifestyle, where these enzyme are regulated according to the presence of nutrients [27]. In connection, many functions involved in signalling and regulation were abundant in the meso- to abyssopelagic zones, which also indicates the presence copiotrophs [31]. In the nutrition depleted deep ocean the carbon sources are thought to be marine snow, carcasses or other nutrition rich lumps, which create a heterogeneous environment with nutrition rich zones. The data suggest that deep marine organisms have developed the ability to move toward and adhere to potential nutrition pellets and otherwise stay in a dormant like state with low metabolic activity [31, 56].

The data analysis also revealed information about phage

and bacteria relations in the samples. Evidence of lytic phages was found in the surface, whereas deeper samples indicated the presence of lysogenic phages. This trend could be caused by the lower cell density generally found in deep ocean and this correlation between cell number, ocean depth and phage strategy in the ocean have been reported elsewhere [14, 54, 57]. Furthermore, phage defence mechanisms were more abundant in deeper samples. Lauro *et al.* [32] suggests the presence of more phage defence systems in copiotrophic bacteria compared to oligotrophic bacteria, which are less susceptible for phage attack due to the slow growth [55]. Hence, this further indicates an enrichment of a copiotrophic lifestyle in the meso- to abyssopelagic zones.

The ratio plot comparing the archipelago samples and southern true ocean samples revealed many OGs unique to the coastal environment. Many of these were of unknown or general functions and this result could be caused by a larger phylogenetic distance between the environments and differential community composition. An interesting observation for the coastal samples was the variety of abundant polysaccharide degrading enzymes, where many of them were acting on molecules connected to algae. The upwelling in coastal areas combined with larger algae blooms could explain the presence of the diverse algae degrading toolbox found near the coast [1]. Furthermore, the coastal upwelling can potentially create areas in the water with low oxygen levels and high presence of nitrogen sources and promote denitrification. This could explain the presences of functions involved in denitrification in the archipelago [13]. Anoxic environments could also promote bacteria involved in sulfur reduction and oxidation [51].

The aphotic part of the ocean is the largest ecosystem on the planet and is dominated by microorganisms. It displays highly similar environmental conditions like stable salinity and temperatures [43]. Looking at the taxonomic composition across the poles, the communities are diverging, however, to a less degree, than through the water column (Hansen *et al.*, *unpublished* [18]). The functions found to differentiate the bathypelagic zones of the Southern Ocean and Arctic Ocean indicates that these differentiations are more related to the taxonomic composition and that the two locations are inseparable on a functional level. More samples would introduce a better statistic ground and higher resolution to compare the deep ocean of the North and South Pole.

4. Materials and Methods

Sequencing and assembly

The arctic ocean water samples were collected and the DNA was purified according to Hansen *et al.*, *unpublished*. The metagenomic library preparation and sequencing were

done according to Kampmann *et al.* [22]. In short, the water samples were obtained and microbes were collected on a 0.2 mm filter after a 2.0-mm pre-filtration. DNA was purified by phenol chloroform extraction and prepared for sequencing using the NEBNext Quick DNA Sample Prep. Master Mix 2 (New England BioLabs Inc., Ipswich, MA, USA). DNA was amplified by PCR and sequenced as 100 bp paired end on an Illumina HiSeq 2000 (Illumina, San Diego, CA, USA). The sequencing adaptors were removed and sequencing reads were trimmed with a custom script (trimming of 10 leading bases, minimum base quality 20, minimum average quality 20). Paired-end and singleton reads were assembled with Meta-IDBA [46] (mink=21, maxk=99, step size of 6) respectively, resulting in two separate assemblies for each metagenome.

Taxonomic annotation

The trimmed reads are aligned to different reference databases, one database after another, using BWA-MEM [33]. Reads were aligned to target organisms with 50% identity at any coverage percentages over 13 base-pairs of mapper scores which were calculated as the read length subtracted by the number of mismatches, insertions and deletions. Target organisms include (1) Microbial complete genomes (NCBI, August 2012) and (2) Microbial draft genomes (NCBI, September 2012). Remaining unmapped reads were mapped against the complete nucleotide database (NCBI, July 2012) using Bowtie2 [30]. If reads mapped to too many organisms, the organism with the biggest mapper-score, i.e. the least mismatches, insertions and deletions, is chosen. However, if various organisms entitle the highest mapper-score for a particular read, only one organism is chosen randomly. The lowest common ancestor on phylum level was selected for visualization.

Non-redundant gene catalogue

MetaGeneMark [64] (default parameters) and Prodigal (codon table 11) [21] were used for gene finding. All predicted genes of all metagenomes were pooled into one bin and homology reduced with CD-HIT-EST (-c 0.95, -n 8, -l 100, -aS 0.9) [35] to create the arctic marine environment gene catalogue. All reads of a metagenome were remapped to the gene catalogue with bwa [34] (default parameters) to create a gene abundance matrix. A mapped read pair to one gene was considered as one observation. If just one mate of a pair mapped to one gene, it was also considered as one observation. However, if each paired-end read mapped to different genes, the mapping was considered as two observations. Mapped singletons were always treated as one observation. Rarefaction steps were calculated as intervals from the remapping process described above. The number of mapped observations populates the abundance matrix with n columns (number of metagenomes) and m rows (number of genes). The

abundance matrix was normalized for the read abundance of the most low abundant metagenome (Equation 1).

$$a_{\text{downsized}} = a * A / A_{\text{min}}$$

$a_{\text{downsized}}$: downsized gene abundance
 a : gene abundance
 A : sample read count
 A_{min} : overall minimal sample read count

(1)

Gene catalogue annotation

All translated genes of the gene catalogue were aligned to the eggNOG database version 4 [39] including all curated orthologous groups (COGs) and non-curated orthologous groups (NOG) and their proteins with BLASTP E-value < 1E-5. The best hits annotation was resumed. The annotation was transferred to the abundance matrix. Genes annotated to the same function were summed up. north vs south: Two samples of the north pole and south were selected according to sequence depth. All functions of the samples were selected from the functional gene catalogue and the mean -log₂ ratio was calculated for each orthologous groups and plotted according to the ratio size. The same ratio was calculated for the 6 south samples and 3 archipelago samples.

Functional analysis

To create an ordination plot of the ocean samples, the normalized COG abundance matrix was log transformed and a Bray-Curtis dissimilarity was calculated between the samples and displayed in an NMDS plot [7]. To observed functional differences throughout the water column, the abundance of the COG functions were used to create a triangular plot of the surface, medium and deep samples. All unknown functions were excluded and a Kruskal-Wallis test between surface, medium and deep samples was conducted on functions that were more than 90% and less than 5% present in surface samples [28]. Based on investigations of the p-value distribution (see Figure S2) a cutoff of 0.001 was chosen for OGs less than 5% present and a cutoff of 0.01 was chosen for samples more than 90% present. Samples of the north (P22 deep and P23 deep) and south (P10 deep and P11 deep) were selected according to sequence depth. All functions of the samples were selected from the functional abundance matrix and the mean -log₂ ratio was calculated for orthologous groups and plotted according to the ratio size. The same was done to infer the differences between the coastal and open ocean environment, the three archipelago samples were compared to the six true open ocean surface, medium and deep samples from the south. The ratio of the mean abundance of the OGs. Two samples of the Arctic Ocean (P10D and P11D) and two samples of the Southern Ocean (P22D and P23D) were selected according to

sequence depth. All functions of the samples were selected from the functional gene catalogue and the mean -log₂-ratio was calculated for each orthologous groups and plotted according to the ratio size. The same ratio was calculated for the six samples of the Southern Ocean (P10D, P11D, P10M, P11M, P10S and P11S) and the three Archipelago samples (P12, P14 and P15).

References

- [1] "Chlorophyll", Dec. 2007.
- [2] Y. H. An and R. J. Friedman, "Concise review of mechanisms of bacterial adhesion to biomaterial surfaces.", *Journal of biomedical materials research*, Vol. 43, No. 3, pp. 338–48, Jan. 1998.
- [3] J. Antonov, D. Seidov, and T. Boyer, World ocean atlas 2009, vol. 2: salinity, U. S. Government Printing Office, Washington D. C., 2006.
- [4] N. Bano, S. Ruffin, B. Ransom, and J. T. Hollibaugh, "Phylogenetic Composition of Arctic Ocean Archaeal Assemblages and Comparison with Antarctic Assemblages", *Applied and Environmental Microbiology*, Vol. 70, No. 2, pp. 781–789, Feb. 2004.
- [5] P. A. Bertram and R. K. Thauer, "Thermodynamics of the Formylmethanofuran Dehydrogenase Reaction in Methanobacterium Thermoautotrophicum", *European Journal of Biochemistry*, Vol. 226, No. 3, pp. 811–818, Dec. 1994.
- [6] E. J. Biers, S. Sun, and E. C. Howard, "Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome.", *Applied and environmental microbiology*, Vol. 75, No. 7, pp. 2221–9, Apr. 2009.
- [7] J. R. Bray and J. T. Curtis, "An Ordination of the Upland Forest Communities of Southern Wisconsin", *Ecological Monographs*, Vol. 27, No. 4, p. 325, Oct. 1957.
- [8] E. D. Chow, O. W. Liu, S. O'Brien, and H. D. Madhani, "Exploration of whole-genome responses of the human AIDS-associated yeast pathogen *Cryptococcus neoformans* var *grubii*: nitric oxide stress and body temperature.", *Current genetics*, Vol. 52, No. 3–4, pp. 137–48, Sept. 2007.
- [9] M. J. Church, H. W. Ducklow, and D. M. Karl, "Light dependence of [3H]leucine incorporation in the oligotrophic North Pacific ocean.", *Applied and environmental microbiology*, Vol. 70, No. 7, pp. 4079–87, July 2004.
- [10] J. R. Cochran, M. H. Edwards, and B. J. Coakley, "Morphology and structure of the Lomonosov Ridge, Arctic Ocean", *Geochemistry, Geophysics, Geosystems*, Vol. 7, No. 5, pp. n/a–n/a, May 2006.
- [11] N. Dasgupta, S. K. Arora, and R. Ramphal, "fleN, a gene that regulates flagellar number in *Pseudomonas aeruginosa*.", *Journal of bacteriology*, Vol. 182, No. 2, pp. 357–64, Jan. 2000.
- [12] E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl, "Community genomics among stratified microbial assemblages in the ocean's interior.", *Science (New York, N.Y.)*, Vol. 311, No. 5760, pp. 496–503, Jan. 2006.

- [13] C. Deutsch, N. Gruber, R. M. Key, J. L. Sarmiento, and A. Ganachaud, "Denitrification and N₂ fixation in the Pacific Ocean", *Global Biogeochemical Cycles*, Vol. 15, No. 2, pp. 483–506, June 2001.
- [14] C. Evans and C. P. D. Brussaard, "Regional variation in lytic and lysogenic viral infection in the Southern Ocean and its contribution to biogeochemical cycling.", *Applied and environmental microbiology*, Vol. 78, No. 18, pp. 6741–8, Sept. 2012.
- [15] S. Freitas, S. Hatosy, J. A. Fuhrman, S. M. Huse, D. B. M. Welch, M. L. Sogin, and A. C. Martiny, "Global distribution and diversity of marine Verrucomicrobia.", *The ISME journal*, Vol. 6, No. 8, pp. 1499–505, Aug. 2012.
- [16] C. G. Friedrich, F. Bardschewsky, D. Rother, A. Quentmeier, and J. Fischer, "Prokaryotic sulfur oxidation", *Current Opinion in Microbiology*, Vol. 8, No. 3, pp. 253–259, 2005.
- [17] N. Fujieda, M. Mori, K. Kano, and T. Ikeda, "Redox properties of quinohemoprotein amine dehydrogenase from *Paracoccus denitrificans*.", *Biochimica et biophysica acta*, Vol. 1647, No. 1-2, pp. 289–96, Apr. 2003.
- [18] J.-F. Ghiglione, P. E. Galand, T. Pommier, C. Pedrós-Alió, E. W. Maas, K. Bakker, S. Bertillon, D. L. Kirchman, C. Lovejoy, P. L. Yager, and A. E. Murray, "Pole-to-pole biogeography of surface and deep marine bacterial communities.", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109, No. 43, pp. 17633–8, Oct. 2012.
- [19] A. M. Goral, K. L. Tkaczuk, M. Chruszcz, O. Kagan, A. Savchenko, and W. Minor, "Crystal structure of a putative isochorismatase hydrolase from *Oleispira antarctica*.", *Journal of structural and functional genomics*, Vol. 13, No. 1, pp. 27–36, Mar. 2012.
- [20] J. T. Hollibaugh, N. Bano, and H. W. Ducklow, "Widespread Distribution in Polar Oceans of a 16S rRNA Gene Sequence with Affinity to Nitrosospira-Like Ammonia-Oxidizing Bacteria", *Applied and Environmental Microbiology*, Vol. 68, No. 3, pp. 1478–1484, Mar. 2002.
- [21] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification.", *BMC bioinformatics*, Vol. 11, No. 1, p. 119, Jan. 2010.
- [22] M.-L. Kampmann, S. L. Fordyce, M. C. Ávila Arcos, M. Rasmussen, E. Willerslev, L. P. Nielsen, and M. T. P. Gilbert, A simple method for the parallel deep sequencing of full influenza A genomes, 2011.
- [23] G. Kelly, S. Prasannan, S. Daniell, K. Fleming, G. Frankel, G. Dougan, I. Connerton, and S. Matthews, "Structure of the cell-adhesion fragment of intimin from enteropathogenic *Escherichia coli*.", *Nature structural biology*, Vol. 6, No. 4, pp. 313–8, Apr. 1999.
- [24] M. B. Klunder, D. Bauch, P. Laan, H. J. W. de Baar, S. van Heuven, and S. Ober, "Dissolved iron in the Arctic shelf seas and surface waters of the central Arctic Ocean: Impact of Arctic river water and ice-melt", *Journal of Geophysical Research*, Vol. 117, No. C1, p. C01027, Jan. 2012.
- [25] M. B. Klunder, P. Laan, R. Middag, H. J. W. de Baar, and K. Bakker, "Dissolved iron in the Arctic Ocean: Important role of hydrothermal sources, shelf input and scavenging removal", *Journal of Geophysical Research*, Vol. 117, No. C4, p. C04014, Apr. 2012.
- [26] K. T. Konstantinidis, J. Braff, D. M. Karl, and E. F. DeLong, "Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre.", *Applied and environmental microbiology*, Vol. 75, No. 16, pp. 5345–55, Aug. 2009.
- [27] K. T. Konstantinidis, J. Braff, D. M. Karl, and E. F. DeLong, "Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre.", *Applied and environmental microbiology*, Vol. 75, No. 16, pp. 5345–55, Aug. 2009.
- [28] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis", *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 583–621, Dec. 1952.
- [29] S. J. Labrie, J. E. Samson, and S. Moineau, "Bacteriophage resistance mechanisms.", *Nature reviews. Microbiology*, Vol. 8, No. 5, pp. 317–27, May 2010.
- [30] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2.", *Nature methods*, Vol. 9, No. 4, pp. 357–9, Apr. 2012.
- [31] F. Lauro and D. Bartlett, "Prokaryotic lifestyles in deep sea habitats", *Extremophiles*, 2008.
- [32] F. M. Lauro, D. McDougald, T. Thomas, T. J. Williams, S. Egan, S. Rice, M. Z. DeMaere, L. Ting, H. Ertan, J. Johnson, S. Ferreira, A. Lapidus, I. Anderson, N. Kyrpides, A. C. Munk, C. Detter, C. S. Han, M. V. Brown, F. T. Robb, S. Kjelleberg, and R. Cavicchioli, "The genomic basis of trophic strategy in marine bacteria.", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 106, No. 37, pp. 15527–33, Sept. 2009.
- [33] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM", Mar. 2013.
- [34] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.", *Bioinformatics*, Vol. 25, No. 14, pp. 1754–60, July 2009.
- [35] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.", *Bioinformatics (Oxford, England)*, Vol. 22, No. 13, pp. 1658–9, July 2006.
- [36] A.-B. Martín-Cuadrado, P. López-García, J.-C. Alba, D. Moreira, L. Monticelli, A. Strittmatter, G. Gottschalk, and F. Rodríguez-Valera, "Metagenomics of the deep Mediterranean, a warm bathypelagic habitat.", *PLoS one*, Vol. 2, No. 9, p. e914, Jan. 2007.
- [37] J. A. Mayfield, C. A. Dehner, and J. L. DuBois, "Recent advances in bacterial heme protein biochemistry.", *Current opinion in chemical biology*, Vol. 15, No. 2, pp. 260–6, Apr. 2011.
- [38] J. R. Mead, S. A. Irvine, and D. P. Ramji, "Lipoprotein lipase: structure, function, regulation, and role in disease.", *Journal of molecular medicine (Berlin, Germany)*, Vol. 80, No. 12, pp. 753–69, Dec. 2002.
- [39] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork, "eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.", *Nucleic acids research*, Vol. 38, No. Database issue, pp. D190–5, Jan. 2010.
- [40] T. Nagata, *Microbial Ecology of the Oceans*, John Wiley & Sons, Inc., d. I. kirc edition, 2008.

- [41] N. Noinaj, M. Guillier, T. J. Barnard, and S. K. Buchanan, "TonB-dependent transporters: regulation, structure, and function.", *Annual review of microbiology*, Vol. 64, pp. 43–60, Jan. 2010.
- [42] H. Ogawa and E. Tanoue, "Dissolved Organic Matter in Oceanic Waters", *Journal of Oceanography*, Vol. 59, No. 2, pp. 129–147, Apr. 2003.
- [43] B. N. Orcutt, J. B. Sylvan, N. J. Knab, and K. J. Edwards, "Microbial ecology of the dark ocean above, at, and below the seafloor.", *Microbiology and molecular biology reviews* : *MMBR*, Vol. 75, No. 2, pp. 361–422, June 2011.
- [44] A. H. Orsi, T. Whitworth, and W. D. Nowlin, "On the meridional extent and fronts of the Antarctic Circumpolar Current", *Deep Sea Research Part I: Oceanographic Research Papers*, Vol. 42, No. 5, pp. 641–673, 1995.
- [45] T. W. Overton, M. C. Justino, Y. Li, J. M. Baptista, A. M. P. Melo, J. A. Cole, and L. M. Saraiva, "Widespread distribution in pathogenic bacteria of di-iron proteins that repair oxidative and nitrosative damage to iron-sulfur centers.", *Journal of bacteriology*, Vol. 190, No. 6, pp. 2004–13, Mar. 2008.
- [46] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data.", *Bioinformatics (Oxford, England)*, Vol. 27, No. 13, pp. i94–101, July 2011.
- [47] J. Pizarro-Cerdá and P. Cossart, "Bacterial Adhesion and Entry into Host Cells", *Cell*, Vol. 124, No. 4, pp. 715–727, 2006.
- [48] K. Porter and Y. Feig, "The use of DAPI for identification and enumeration of bacteria and blue-green algae", *Limnology and oceanography*, Vol. 25, p. 943, 1980.
- [49] D. Rusch, A. Halpern, and G. Sutton, "The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific", *PLoS biology*, 2007.
- [50] E. Schwartz, A. Henne, R. Cramm, T. Eitinger, B. Friedrich, and G. Gottschalk, "Complete Nucleotide Sequence of pHG1: A *Ralstonia eutropha* H16 Megaplasmid Encoding Key Enzymes of H₂-based Lithoautotrophy and Anaerobiosis", *Journal of Molecular Biology*, Vol. 332, No. 2, pp. 369–383, 2003.
- [51] S. M. Sievert, R. P. Kiene, and H. N. Schultz-Vogt, "The sulfur cycle", June 2007.
- [52] T. A. Springer, "Folding of the N-terminal, ligand-binding region of integrin α -subunits into a α -propeller domain", *Proceedings of the National Academy of Sciences*, Vol. 94, No. 1, pp. 65–72, Jan. 1997.
- [53] C. Stephens, J. Antonov, and T. Boyer, World Ocean Atlas 2001. Volume 1, Temperature, U. S. Government Printing Office, Washington D. C., 2002.
- [54] F. M. Stewart and B. R. Levin, "The population biology of bacterial viruses: why be temperate.", *Theoretical population biology*, Vol. 26, No. 1, pp. 93–117, Aug. 1984.
- [55] C. A. Suttle, "Marine viruses—major players in the global ecosystem.", *Nature reviews. Microbiology*, Vol. 5, No. 10, pp. 801–12, Oct. 2007.
- [56] P. Thureborn, D. Lundin, J. Plathan, A. M. Poole, B.-M. Sjöberg, and S. Sjöling, "A metagenomics transect into the deepest point of the baltic sea reveals clear stratification of microbial functional capacities.", *PloS one*, Vol. 8, No. 9, p. e74983, Jan. 2013.
- [57] M. Weinbauer, I. Brettar, and M. Höfle, "Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters", *Limnology and Oceanography*, 2003.
- [58] W. B. Whitman, "Prokaryotes: The unseen majority", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 95, No. 12, pp. 6578–6583, June 1998.
- [59] C. J. Whittaker, C. M. Klier, and P. E. Kolenbrander, "Mechanisms of adhesion by oral bacteria.", *Annual review of microbiology*, Vol. 50, pp. 513–52, Jan. 1996.
- [60] H. Wiggerich, B. Klauke, R. Koplin, U. Priefer, and A. Puhler, "Unusual structure of the tonB-exb DNA region of *Xanthomonas campestris* pv. *campestris*: tonB, exbB, and exbD1 are essential for ferric iron uptake, but exbD2 is not", *J. Bacteriol.*, Vol. 179, No. 22, pp. 7103–7110, Nov. 1997.
- [61] L. Worthington, "Genesis and evolution of water masses", *Woods Hole Oceanographic Institution*, 1968.
- [62] J. Wu, W. Gao, R. Johnson, W. Zhang, and D. Meldrum, "Integrated Metagenomic and Metatranscriptomic Analyses of Microbial Communities in the Meso-and Bathypelagic Realm of North Pacific Ocean", *Marine drugs*, 2013.
- [63] J. A. Yethon and C. Whitfield, "Purification and characterization of WaaP from *Escherichia coli*, a lipopolysaccharide kinase essential for outer membrane stability.", *The Journal of biological chemistry*, Vol. 276, No. 8, pp. 5498–504, Feb. 2001.
- [64] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences.", *Nucleic acids research*, Vol. 38, No. 12, p. e132, July 2010.

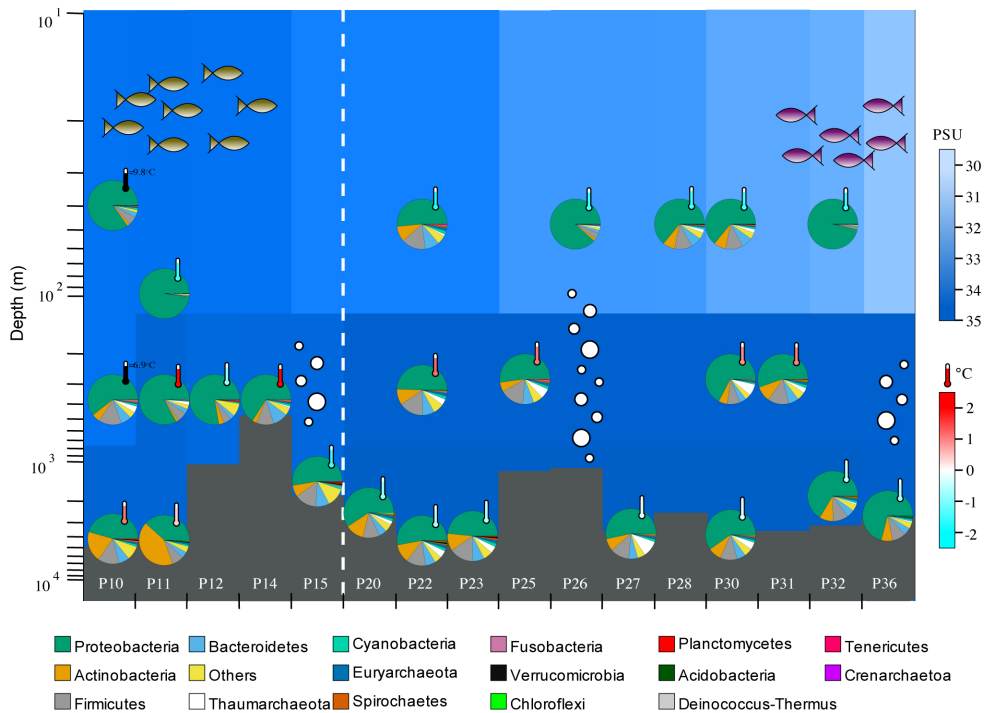


Figure 1: The figure represents metadata and taxonomic distribution of 25 water samples collected during the Galathea III and LOMROG II polar expeditions. The taxonomic distribution of each sample is indicated as pie-charts together with the temperature. Proteobacteria are the most abundant phyla across the samples. All sample temperatures range from -2°C to 2°C , except the surface and medium samples of location P10 (7°C and 9°C respectively). The positioning of the pie-charts represents the sampling depth. The blue background coloring scheme represents salinity ranging from 34 PSU to 35 PSU in most samples. The salinity of the arctic surface samples displayed concentrations between 31 PSU and 33.5 PSU.

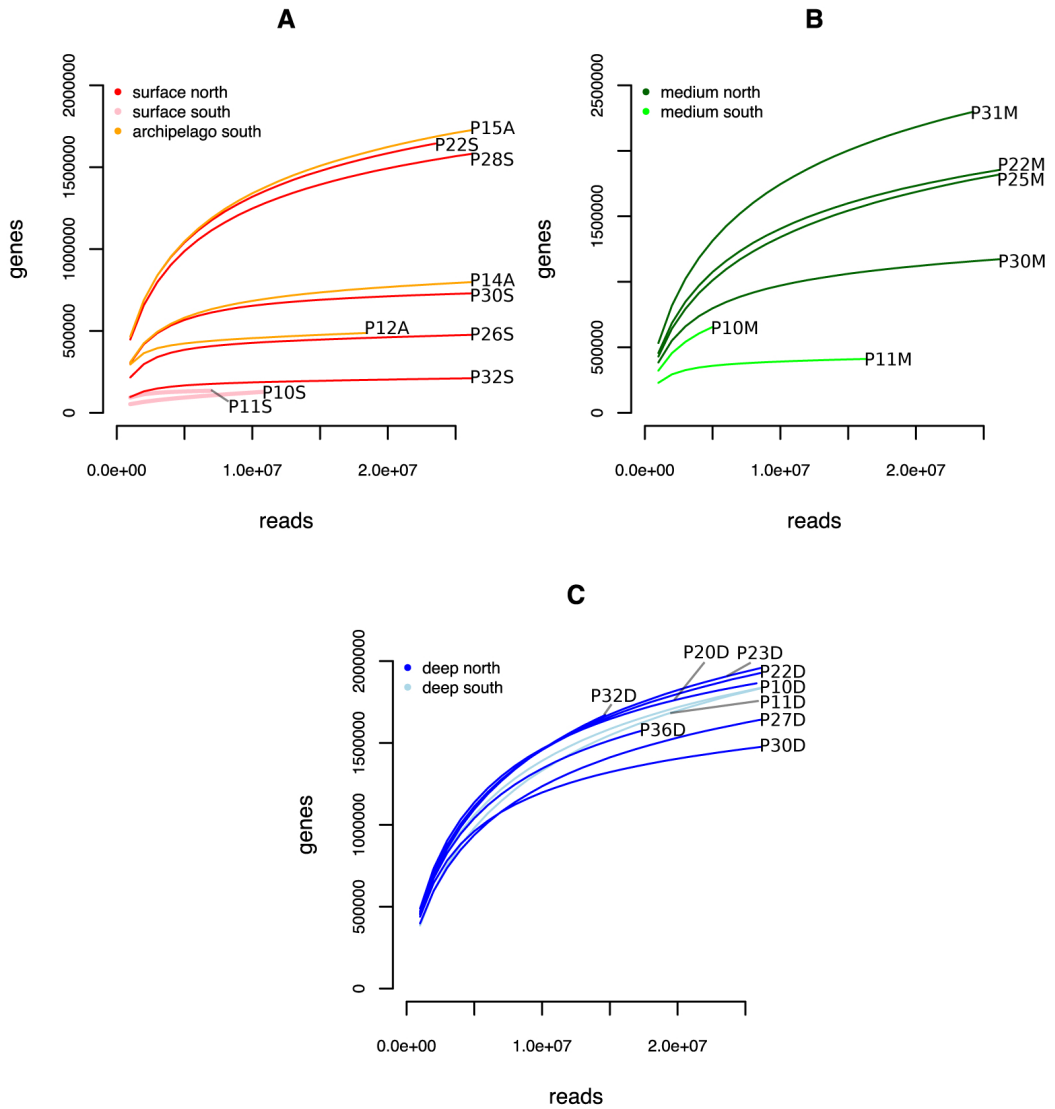


Figure 2: Rarefaction curves for surface (A), medium (B) and deep (C) samples, reads were remapped to the non-redundant gene catalogue.

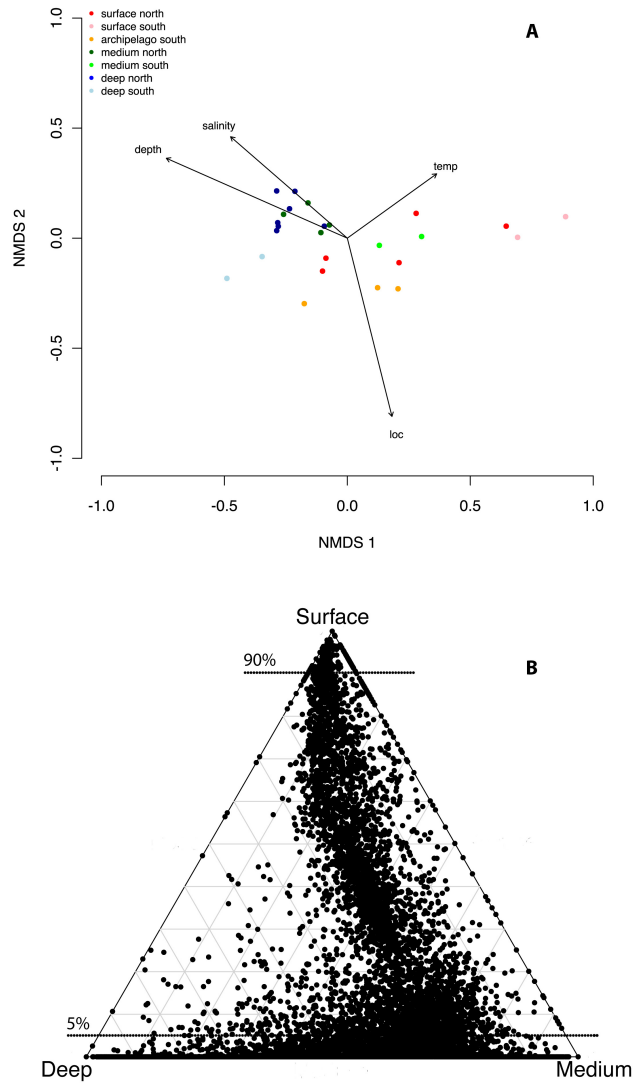


Figure 3: The figure presents the functional clustering of surface, medium, deep and archipelago samples. The NMDS plot displays functional correlation of all 25 samples (A). The triangular displays the relative abundance of functional annotation of genes annotated to eggNOG database (B). A cutoff of 5% and 90% relative to surface were chosen for further analysis.

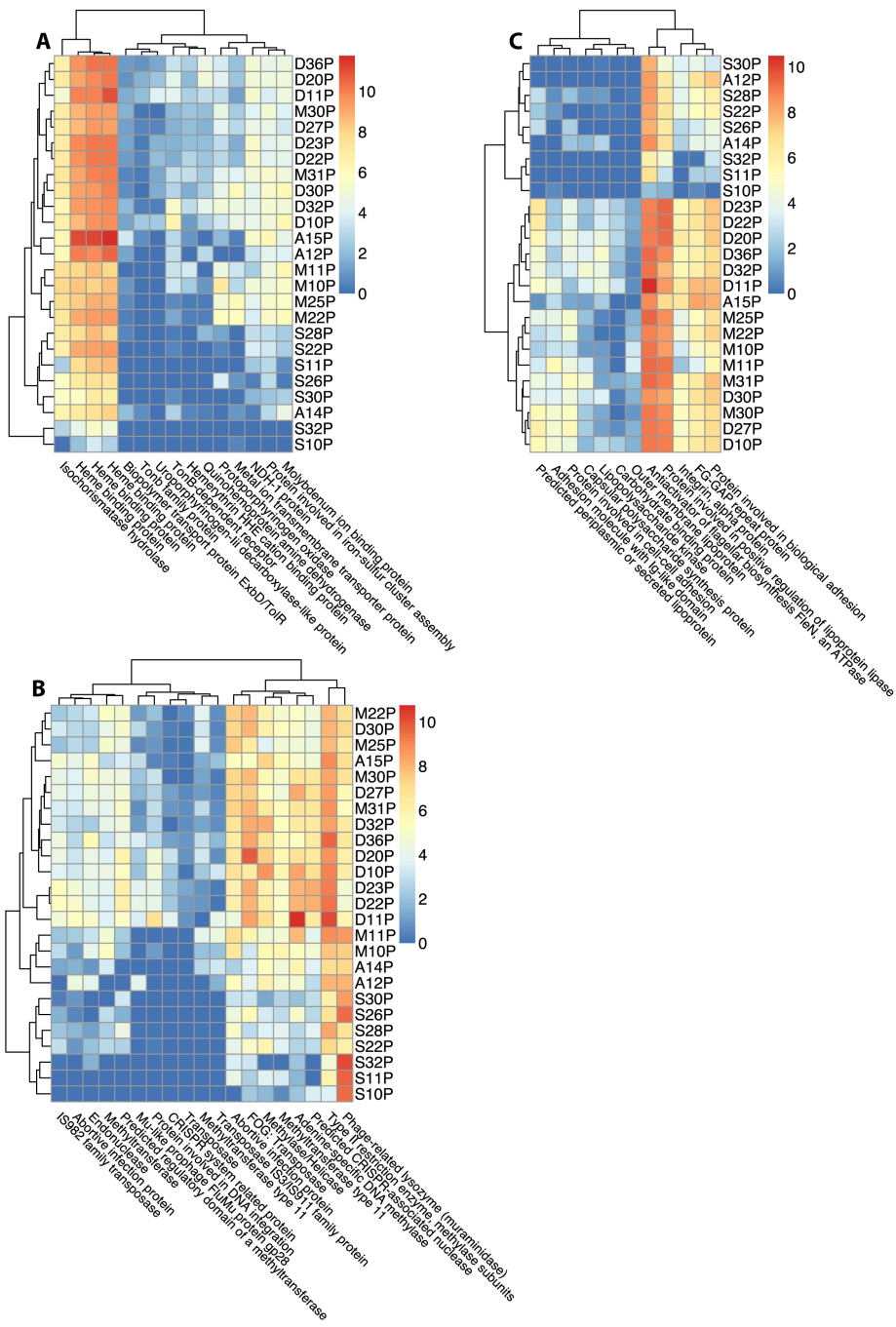


Figure 4: Heatmap displaying OG abundance for three functional categories; A: Iron and other metals, B: phages, C: adhesion and motility. OGs were clustered with Euclidean distance (A and B) and Manhattan distance (C).

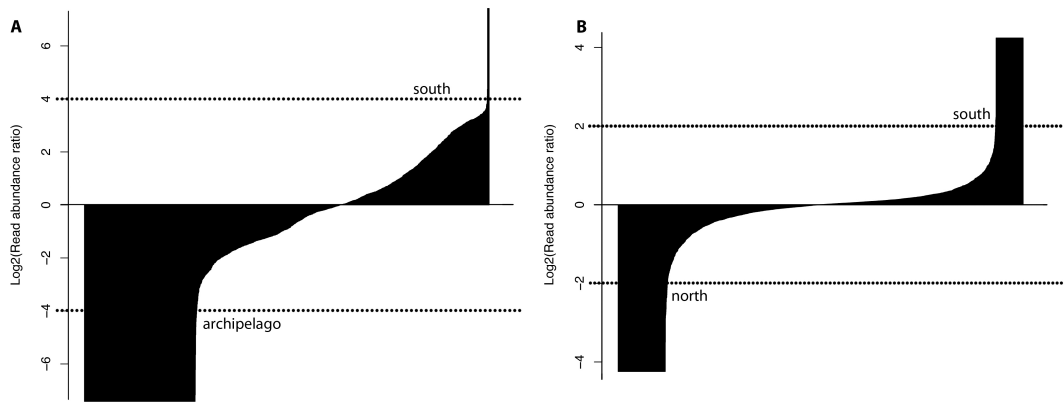


Figure 5: Log₂ mean abundance ratio of archipelago versus Southern Ocean samples (A) and Arctic versus Southern Ocean samples (B). OGs with log-ratios >4 were classified as being higher abundant in true ocean and the OGs with log-ratios <-4 were classified as being higher abundant in coastal waters. OGs with a log-ratio >2 were classified as abundant in deep South Pole samples and log-ratio <-2 were classified as abundant in deep North Pole samples.

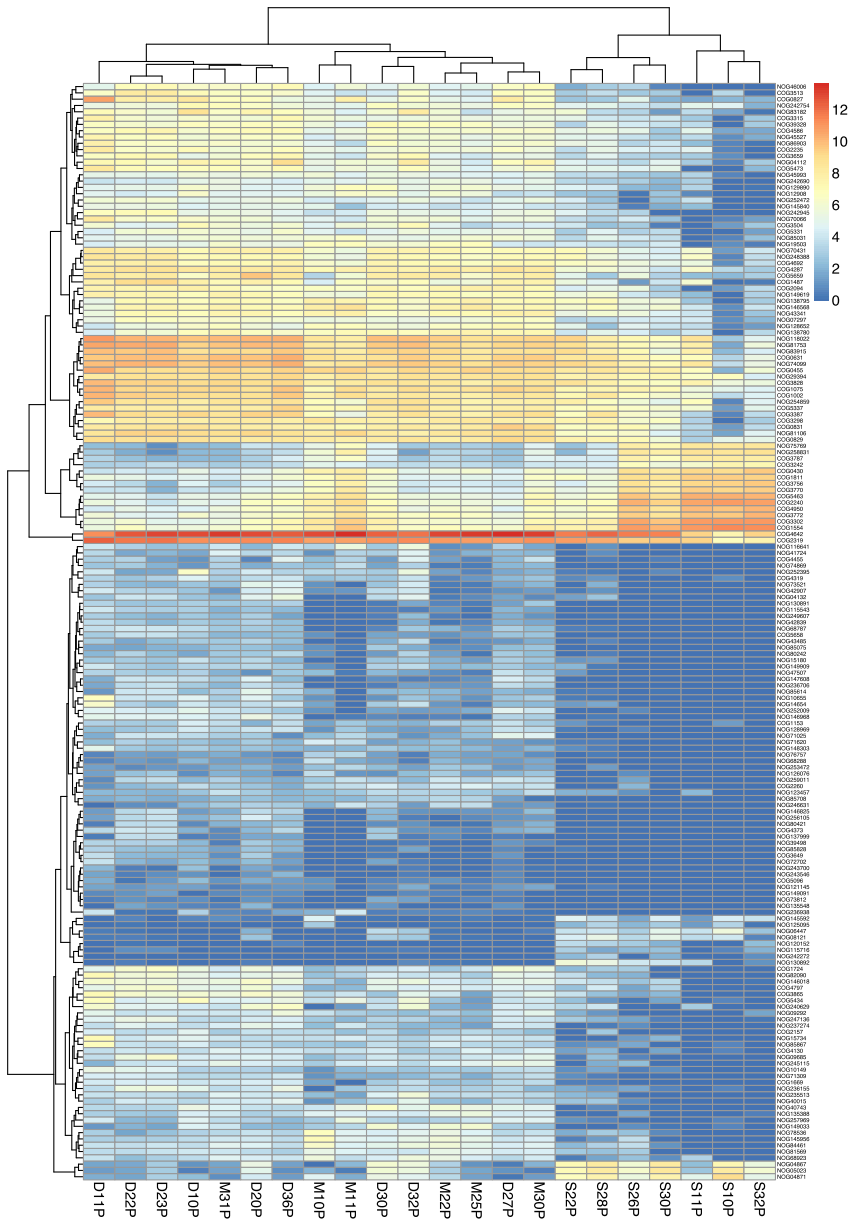


Figure S1: Heatmap displaying the OG abundance. 176 diverse functions were identified, the functions' eggNOG identifiers are displayed. 26 OGs showed higher abundance in surface samples and 150 OGs displayed higher abundance in the medium and deep layers. Clusters of OGs that preferred either the medium or deep environment were not detected.

Table S1: Metadata of the 25 polar metagenomic samples collected during the Galathea III expedition in the Southern Ocean and LOMROG II expedition in the Arctic Ocean.

Sample	Location	Type	Sampling depth [m]	Total depth [m]	Temperature [°C]	Salinity [PSU]	Volume [L]	Latitude	Longitude
A12P	South	Archipelago	400	1,116	-0.738	34.4783	90	-63.54122	-61.38004
A14P	South	Archipelago	400	500	1.983	34.5411	60	-62.57909	-58.03619
A15P	South	Archipelago	1,500	1,956	-1.578	34.5411	330	-62.18498	-57.45811
D10P	South	Deep	4,200	5,376	1.045	34.7071	240	-53.17727	-173.02781
D11P	South	Deep	4,200	4,568	0.361	34.6972	240	-66.35371	-108.55849
D20P	North	Deep	2,500	2,650	-0.78	34.921	77	81.5645	16.4074
D22P	North	Deep	4,300	4,460	-0.638	34.941	105	88.1284	59.4161
D23P	North	Deep	4,300	4,457	0.638	34.941	168	88.1615	65.3502
D27P	North	Deep	3,850	3,912	-0.272	34.953	112	88.088	157.0739
D30P	North	Deep	3,885	3,934	-0.268	34.953	112	88.3581	-175.3815
D32P	North	Deep	2,000	3,490	-0.494	34.931	84	88.2713	-129.2037
D36P	North	Deep	2,800	2,863	-0.739	34.926	140	88.4374	-58.2728
M10P	South	Medium	400	5,376	6.891	34.3421	60	-53.17727	-173.02781
M11P	South	Medium	400	4,568	1.983	34.6543	60	-66.35371	-108.55849
M22P	North	Medium	350	4,460	1.17	34.859	28	88.1284	59.4161
M25P	North	Medium	300	1,258	1.068	34.848	70	88.3292	133.1607
M30P	North	Medium	300	3,934	0.873	34.838	28	88.3581	-175.3815
M31P	North	Medium	300	3,898	0.978	34.845	42	88.5069	-156.1645
S10P	South	Surface	40	5,376	9.811	34.3698	60	-53.17727	-173.02781
S11P	South	Surface	100	4,568	-1.568	34.1226	60	-66.35371	-108.55849
S22P	North	Surface	50	4,460	-1.802	33.515	28	88.1284	59.4161
S26P	North	Surface	50	1,198	-1.544	32.596	105	88.2465	150.0963
S28P	North	Surface	50	2,725	-1.507	32.565	49	88.4266	158.5342
S30P	North	Surface	50	3,934	-1.481	32.018	28	88.3581	-175.3815
S32P	North	Surface	50	3,490	-1.484	31.783	84	88.2713	-129.2037

Table S1 continued: Assembly statistics and number of genes called from the metagenomic samples collected during the Galathea III expedition in the Southern Ocean and LOMROG II expedition in the Arctic Ocean.

Sample	Number of bp	bp after trimming	% bp used	N50	# Contigs	# Genes MetaGeneMark	# Genes Prodigal
A12P	6,752,312,700	5,370,616,035	0.80	2410	12407	16264	9351
A14P	14,107,389,100	10,463,542,337	0.74	864	116172	132094	58636
A15P	22,875,763,680	17,276,666,926	0.76	1419	776801	1061644	595502
D10P	19,242,853,504	13,918,759,303	0.72	936	441113	520346	251312
D11P	38,990,413,280	37,713,389,214	0.97	957	591350	1417200	697542
D20P	11,963,483,868	8,350,865,051	0.70	1070	246335	294250	152218
D22P	13,211,062,872	8,612,117,894	0.65	1044	369017	460604	227579
D23P	27,619,036,592	14,181,863,965	0.51	1009	580522	694168	350357
D27P	36,173,321,156	24,488,207,580	0.68	1006	355749	421894	210799
D30P	15,804,345,536	11,074,851,937	0.70	1010	162273	188181	94492
D32P	9,953,652,632	4,931,200,759	0.50	798	185901	194363	87265
D36P	9,441,133,996	4,772,110,910	0.51	839	167400	188501	82162
M10P	3,702,547,996	2,422,291,751	0.65	892	47965	54185	25076
M11P	5,772,477,144	4,559,972,170	0.79	1274	24381	30857	15622
M22P	15,518,491,884	10,243,146,369	0.66	774	459212	481995	219583
M25P	16,606,058,252	10,373,370,228	0.62	957	453369	530494	256746
M30P	10,728,299,708	7,586,418,205	0.71	1233	97178	126105	65354
M31P	14,667,415,588	7,816,420,744	0.53	865	359299	402010	188463
S10P	5,806,809,288	4,468,634,960	0.77	2276	10613	12472	6213
S11P	1,794,302,384	1,381,477,855	0.77	12422	1069	5473	4620
S22P	11,254,647,320	7,862,736,245	0.70	781	455895	491061	219735
S26P	8,459,648,316	6,430,473,986	0.76	4815	9739	10822	6088
S28P	18,308,070,116	12,604,143,340	0.69	1238	503378	652496	351093
S30P	11,280,164,560	7,838,826,568	0.69	821	59386	115692	49790
S32P	10,813,050,500	8,004,356,894	0.74	40444	2022	5305	4489

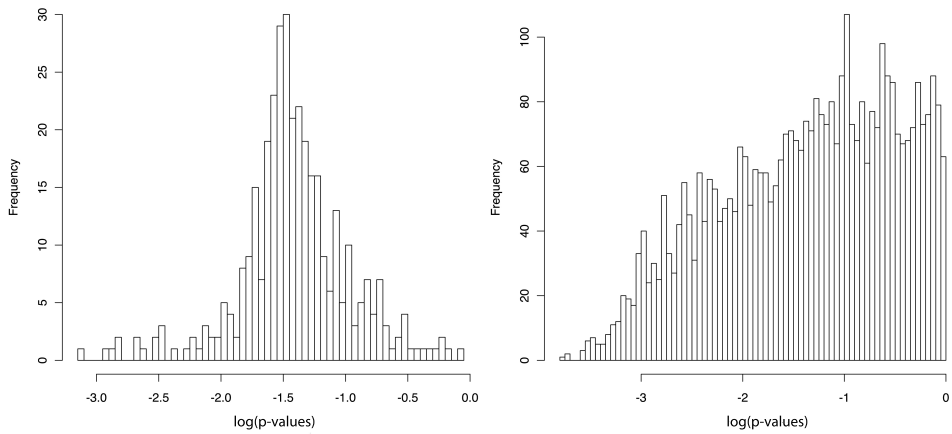


Figure S2: P-value distribution of Kruskal-Wallis test between surface, medium and deep samples on functions that were more than 90% (left) and less than 5% (right) present in surface samples. A p-value cutoff of 0.001 was chosen for OGs less than 5% present and a cutoff of 0.01 was chosen for samples more than 90% present.

Chapter 10

Manuscript III

10.1 Proteolytic enzymes of the polar marine environment - Patent application

Psychrotrophic bacteria are known to produce various proteases that differ in optimal pH and temperature compared to proteases isolated from moderate environments [93]. The arctic environment is a unique habitat, where the bacterial flora has been adapted to grow at low temperatures. However, few research articles have been published on proteases from the arctic marine environment [36, 108, 170, 189]. Exploiting the natural resources of the polar marine environment still has a high potential for bioprospecting and identification of novel proteases. Proteases with their annual sale worth of 1.5 – 1.8 billion US dollars are a valuable product for the industry [76, 176]. They find applications in industries such as leather manufacturing, food processing, detergents, pharmaceuticals and bioremediation (described more extensively in Chapter 4).

The industrial demand for novel enzymes prompted us to search for proteases, which have adapted to the extreme environments of the polar oceans. Here we describe a comprehensive *in silico* metagenomics screen where we analyze 26 metagenomes from the polar marine environment sampled at depths between 40 m and 4300 m. The presented bioinformatic study aims to find proteases with the focus on identifying novel candidate sequences. We compare HMM and alignment based protease identification methods for dataset screening. We present the distribution of proteolytic enzymes across all families and subfamilies, which might be potential candidates for expression trials. 2,707 novel protease sequences were identified that cannot be found in public databases or metagenomes, including the Global Ocean Sampling Expedition [140].

This work provides a pivotal step toward identification of proteases with

novel catalytic activities from polar marine environments. We anticipate that our findings can bridge exploratory science with novel biotechnological processes and innovations. Due to the commercial potential of the unique dataset, we filed a patent application at DTU (the patent application is appended later in the chapter). Negotiations between DTU and interested industrial partners are still ongoing to reach an agreement of the data usage. Thus, no sequence information has been provided and the manuscript will not be published in the near future to avoid public access to raw data.

Anmeldelse af opfindelse gjort ved DTU

Hent altid anmeldesskemaet fra DTU Portalen. Så er du sikker på, at du bruger den nyeste version.

Anmeldesskemaet udfyldes og underskrives. Husk at vende opfindelsen med den innovationsansvarlige eller institutdirektøren.

Skemaet sendes elektronisk til email-adressen: patentadm@adm.dtu.dk både som indskannet dokumentet med underskrifter og som Microsoft Word-fil uden underskrifter.

Har du spørgsmål om, hvorledes du udfylder arket, så kontakt dit instituts innovationsansvarlige eller søg på 'Patents' på DTU Portalen.


Titel på opfindelse

Proteolytic enzymes of the polar marine environment

Interne DTU-opfindere

(Der er plads til yderligere opfindere til slut i dokumentet. Husk at det kun er DTU-ansatte, som skal anmelde på dette skema. I forbindelse med indlevering af patentansøgning på opfindelser til de amerikanske og internationale patentmyndigheder er det et krav, at man oplyser opfindernes privatadresser og nationaliteter. Disse oplysninger bedes derfor anført nedenfor. Ydermere skal DTU orienteres om eventuelle adresseændringer)

DTU Opfinder 1 (Kontaktperson)	Institut:	DTU Biosustain	Andel af opfindelsen (%)	25
Fulde navn:	Nikolaj Blom		Stilling:	Seniorforsker
Email (arbejde):	blom@cbs.dtu.dk		Tlf. (arbejde):	[REDACTED]
Nationalitet:	Dansk		Dato og underskrift:	6/8-13 <i>N. Blom</i>
Privat adresse: <small>(newline: Shift+Enter)</small>	[REDACTED]			
DTU Opfinder 2	Institut:	DTU Systembiologi	Andel af opfindelsen (%)	25
Fulde navn:	Thomas Sicheritz-Pontén		Stilling:	Professor
Email (arbejde):	thomas@cbs.dtu.dk		Tlf. (arbejde):	[REDACTED]
Nationalitet:	Østrigisk		Dato og underskrift:	<i>Dr. Thomas Sicheritz-Pontén</i>
Privat adresse:	[REDACTED]			
DTU Opfinder 3	Institut:	DTU Systembiologi	Andel af opfindelsen (%)	25
Fulde navn:	Josef Korbinian Vogt		Stilling:	Phd studerende
Email (arbejde):	josef@cbs.dtu.dk		Tlf. (arbejde):	[REDACTED]
Nationalitet:	Tysk		Dato og underskrift:	<i>Josef K Vogt</i>
Privat adresse:	[REDACTED]			

DTU Opfinder 4	Institut:	DTU Biosustain	Andel af opfindelsen (%)	25
Fulde navn:	Henrik Marcus Geertz-Hansen		Stilling:	PhD studerende
Email (arbejde):	hmgh@cbs.dtu.dk		Tlf. (arbejde):	
Nationalitet:	Dansk		Dato og underskrift:	
Privat adresse:				
DTU Opfinder 5	Institut:		Andel af opfindelsen (%)	
Fulde navn:			Stilling:	
Email (arbejde):			Tlf. (arbejde):	
Nationalitet:			Dato og underskrift:	
Privat adresse:				
DTU Opfinder 6	Institut:		Andel af opfindelsen (%)	
Fulde navn:			Stilling:	
Email (arbejde):			Tlf. (arbejde):	
Nationalitet:			Dato og underskrift:	
Privat adresse:				

Ekstern opfinder 1	Andel af opfindelsen (%)	
Fulde navn:	Stilling:	
Email (arbejde):	Tlf. (arbejde):	
Nationalitet:	Privat adresse:	
Ekstern opfinder 2	Andel af opfindelsen (%)	
Fulde navn:	Stilling:	
Email (arbejde):	Tlf. (arbejde):	
Nationalitet:	Privat adresse:	
Ekstern opfinder 3	Andel af opfindelsen (%)	
Fulde navn:	Stilling:	
Email (arbejde):	Tlf. (arbejde):	
Nationalitet:	Privat adresse:	

Underskrift Institutsdirektør/Patentansvarlig			
Dato		Underskrift	

Teknologien

(I denne sektion skal opfindelsen beskrives sammen med udviklingsstadiet)

Beskriv opfindelsen

(Beskriv opfindelsens hovedtræk og det teknologiområde, opfindelsen angår, på en alment forståelig måde. Kopier gerne figurer og grafer ind, vedhæft gerne yderligere materiale og udvid skrivefeltet hvis nødvendigt)

The invention consists of novel dna sequences encoding peptide-cleaving enzymes, proteases, that function at low temperature (below 25 C).

The sequences were obtained under the research project "DNA of the Polar Seas". This includes water samples collected at two Danish marine research expeditions: The 2006-2007 Galathea3 expedition and the 2009 LOMROG-II expedition. The samples were collected in international waters close to either Antarctica or the geographic North Pole.

Large volumes of water (50-300 liters) were filtered for microorganisms (between 0.2 – 2.0 microns in size), dna was extracted and sequenced using a shot-gun metagenome approach.

Sequence reads were assembled

Beskriv udviklingsstadiet

(Er der f.eks. en fungerende prototype, er nogle delelementer blevet testet, eller hvilke indikationer er der på, at dette ville kunne fungere)

Beskriv det kommende års forventede resultater inklusiv forventede udfordringer

(Hvilket forsknings- og udviklingsarbejde forventes at gøres som styrker opfindelsen teknisk eller kommercielt, og hvor ser I de største udfordringer?)

In the coming year we expect to team up with a strong experimental partner that can test some of the identified proteases in relevant assays. This will require expertise in gene cloning, expression and purification and running assays testing the function of the enzymes at various temperatures.

The biggest challenge will be to find a partner where all of the expertise mentioned above is present and running routinely. If a single, strong partner is not identified, it will be difficult to manage testing as the inventors have little activity in the laboratory.

**10.2 Exploiting the polar marine environment for
bioprospecting: novel protease discovery**

Exploiting the polar marine environment for bioprospecting: novel protease discovery

Josef Korbinian Vogt¹, Henrik Marcus Geertz-Hansen^{1,2,3},
Lea Benedicte Skov Hansen⁴, Søren Sørensen⁴, Jesper Salomon³,
Thomas Sicheritz-Pontén^{1,2} and Nikolaj Blom^{1,2}

¹Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark

²Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark,

Kogle Alle 6, DK-2970 Hørsholm, Denmark

³Novozymes A/S, Krøgshøjvej 36, DK-2880 Bagsværd, Denmark

⁴Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark

ABSTRACT – Proteases with an annual sale worth of 1.5 – 1.8 billion US dollars are a valuable resource for the industry. They find applications in industries such as leather manufacturing, food processing, detergents, pharmaceuticals and bioremediation.

The industrial demand for novel enzymes prompted us to search for proteases, which have adapted to the extreme environments of the polar oceans. Here we describe a comprehensive *in silico* metagenomics screen where we analyze 26 metagenomes from the polar marine environment sampled at depths between 40 m and 4300 m. We present the distribution of proteolytic enzymes across all families and subfamilies, which might be potential candidates for expression trials. This work provided a pivotal step toward identification of proteases with novel catalytic activities from polar marine environments. We anticipate that our findings can bridge exploratory science with novel biotechnological processes and innovations.

KEY WORDS – metagenomics, enzymes, proteases, bioprospecting, deep-sea, polar oceans

1. Introduction

Microorganisms are essential in today's efforts to produce secondary metabolites [18, 26] and enzymes [5, 8]. In this context the term bioprospecting has been coined for the systematic search for these products in environmental samples.

In the search for new enzymes, much effort has been directed towards extremophiles. These microorganisms inhabit environments characterized by extreme physical or chemical conditions and consequently have evolved enzymes with correspondingly extreme properties [5].

The arctic deep-sea environment can be described as extreme due to its obvious characteristics, such as constant low temperature, depletion of light and almost famine conditions. This habitat is a fertile ground for bioprospecting as its natural resources are abundant and have not been fully exploited for extremophilic enzymes. The industrial applicability and high value of proteases has increased the demand for discovery of proteases more adapted to the specific conditions of particular industrial processes.

This study focuses on proteolytic enzymes (proteases), which represent one of the most diverse enzyme classes with an estimated annual sale worth of 1.5 – 1.8 billion

US dollars [12, 31]. Proteases find industrial applications within leather manufacturing, food processing, pharmaceuticals, detergents and bioremediation [1, 3, 10, 16, 19]. Detergent proteases, with an annual market of about 1 billion US dollars, account for the largest protease application segment [31]. Identification and expression of proteases from extreme environments have been reported [21, 29, 32]. There has only been one report of a marine metagenome derived protease, which involved isolation and characterization of a metalloprotease from deep-sea sediment metagenomic libraries [15]. To our knowledge no extensive *in silico* protease mining of polar marine (deep-sea) environments has been conducted to date. Here we present a workflow for identifying protease sequences across all families and subfamilies from metagenomic data. 26 metagenomes were obtained from environmental samples collected during the Galathea III and LOMROG II expeditions at a depth range of 40 – 4,300m. The temperatures within those regions range from -2°C to 2°C. However, the medium sample taken north of the Antarctic Circumpolar Current (sample 10M) displays a warmer environment, with a measured temperature of 7°C. The concentrations were relatively stable between 34 PSU and 35 PSU, except in the surface samples from the North

Pole, which ranged from 31 to 33.5 PSU. Within those extreme environmental samples we identified 2,707 novel protease candidates, which cannot be found in public databases or metagenomic datasets.

2. Results

The study compasses 26 metagenomes, which were collected during the polar marine expeditions Galathea III and LOMROG II. Water samples at 16 different locations and also at varying depths were collected.

The assembled metagenomes were scanned with the two gene prediction algorithms Prodigal and MetaGeneMark. Prodigal is optimized for microbial gene calling, whereas MetaGeneMark is designed for gene calling in metagenomes and novel prokaryotes. Combining the results of the two gene prediction algorithms yields a total of 13,919,017 predicted coding sequences in the 26 metagenomes. To quantify the gene contribution of each gene caller for each sample, the called genes were combined and clustered using CD-HIT (Figure 1). The sampled specific clustered gene count (Figure 1 blue bar) exceeds the number of genes called by Prodigal alone (Figure 1 red bar), but is generally lower than the number of genes called by MetaGeneMark. To remove gene redundancy between samples, the gene catalogue of >13million genes was homology reduced using CD-HIT. The non-redundant polar marine metagenome gene catalogue consists of 5,218,92 genes in total when the longest gene of a cluster was used as representative. In this catalogue, 972,738 genes (18,6%) were called by Prodigal and 4,246,189 genes (81,4%) were called by MetaGeneMark.

The translated genes were aligned to Swiss-Prot using BLASTp and the annotation of the best hit was resumed. 626,257 translated genes can be annotated to proteins in the Swiss-Prot database (Figure S1) out of which 16,278 (2,6%) are acting on peptide bonds (Figure 2A, plots were created using Krona [22]). Metallo, serine and cysteine family proteases are most abundant in the non-redundant gene catalogue.

For targeted protease discovery, a proteases specific search workflow was applied; the non-redundant gene catalogue was searched with protease specific hidden Markov models (HMMs). A detailed procedure for creating the HMMs is provided in the methods section. The search led to the identification of 37,861 potential proteases of which 19,328 sequences (51%) originate from deep-sea samples, 8,694 sequences (23%) from medium depth samples and 9,839 sequences (26%) from surface samples. The HMM and BLASTp annotated proteases share a sequence overlap of 12,350 sequences. The HMM workflow is able to pick up most alignment based annotated sequences but

also identifies potential targets which are not part of Swiss-Prot. To identify potential protease sequences that are unique to the polar marine environment samples collected for this study, the 37,861 candidate sequences were aligned to UniProt and assembled metagenomes obtained from CAMERA using BLASTp and tBLASTn. The parsing parameters were adjusted so that even weak sequence similarities were captured. This led to the removal of 35,154 sequences (92.8%) due to homology to the query databases and or datasets. A detailed overview of the aligned protease sequences is provided in Table 1.

2,707 (7.2%) protease sequences can only be found in the polar marine gene catalogue. Out of these proteases, 1,531 sequences originate from deep-sea samples, 540 from medium depth samples and 636 from surface samples (Figure S2). After novelty enrichment we find that the percentage of proteases originating from deep-sea samples slightly increases from 51% to 56,6%. We were able to identify 47 sequences belonging to the A protease family sequences, 137 C protease family, 965 M family, 1325 to the S family, 203 to the T protease family and 28 to the U protease family (Figure S2). A detailed overview of the subfamily distribution is provided in Figure S3.

In order to get an understanding of the sequence identity between the identified protease sequences, the 37,861 protease sequences before novelty enrichment and the 2,707 protease sequences after novelty enrichment were homology clustered according to MEROPS families and the 10 most abundant MEROPS subfamilies.

The cluster counts drop rapidly when the cluster threshold is decreased before novelty enrichment (Figure 3 A and C). However, after novelty enrichment, the cluster count decreases slower when decreasing the sequence identity threshold for clustering (Figure 3 B and D). This indicates that novelty enrichment does not just remove sequences, which are present in other environments and databases, but also decreases the relative inter-sequence diversity.

Industrial enzymes are preferably secreted from a heterologous production organism. This way, the enzyme is released to the medium and cell disruption can be avoided which makes the downstream processing less complicated. In order to narrow down the list of candidates for potential industrial applications, the 2,707 sequences were scanned for signal peptides via SignalP 4.1. 687 sequences have a signal peptide, indicating extracellular proteases.

3. Discussion

In silico mining for proteases in metagenomes has not been reported as extensively as the study we present. In general there are only few reports of proteases derived from metagenomic studies available [14]. Most bioprospecting efforts or screening for novel enzymes are driven by functional metagenomics [5, 6, 20, 34] and do not fully

exploit the potential of sample sequencing. With functional metagenomic approaches one might not be able to express enzymes, which may be valuable for industrial processes. The presented workflow does not depend on prior protein expression for identifying target sequences and makes full use of the sequencing data by assembly and gene prediction. Already targeted sequences are more likely to be successfully expressed in an expression host. The HMM approach, which incorporates Novozymes internal database of characterized proteins, makes identification of proteases superior over alignment based annotation via BLAST. The HMM based protease finding yields 25,511 potential protease sequences more than a BLAST based approach. With our workflow we are able to find 12,350 sequences, which were also found by a BLAST alignment. However, the HMMs were not able to annotate 3,296 sequences, which the BLAST approach was able to identify. Furthermore, novelty enrichment ensures that these particular sequences have no known close homologues, neither in the public database UniProt nor in other metagenomes, including other marine metagenomes.

The polar marine environment is a resource, which has not fully been exploited for extremophilic proteases. To date, only one report of a marine metagenome-derived protease exists which is based on the isolation and characterization of a metalloprotease from a deep-sea sediment metagenomic library [15]. In our study we identified over 2,000 protease sequences of polar marine samples of depths up to 4,300 m. To our knowledge no specific *in silico* studies have been conducted to search for proteases in polar marine metagenomes of such depths. Applications of extremophilic proteases from the polar marine metagenomes could be high catalytically efficient enzymes at low temperature suitable for the detergent market. The detailed mapping on protease subfamily level presented here enabled us to identify a large variety of proteases including serine proteases, which may be important for the detergent industry.

Extracellular proteases catalyze the hydrolysis of large proteins to smaller molecules for subsequent absorption by the cell. The objective of cloning bacterial protease genes has mainly been the overproduction of enzymes for various commercial applications in food, detergent or pharmaceutical industries [27]. We were able to identify 687 of such protease sequences with an intrinsic signal peptide sequence. Excreted proteins are preferred targets for industrial scale expression as issues such as intracellular aggregation inclusion bodies can be avoided. However, intracellular proteases may still be of industrial interest as heterologous expression can be achieved through the design of fusion proteins or addition of heterologous signal peptides [9].

We anticipate that expression trials of sequences identified in our study would contribute to industrial processes which benefit from the adaptation of microbes to the polar marine environment. With our approach we are able

to analyze the metagenomes in a resource efficient way to screen for potential industrial targets. The number of sequences can be narrowed to a manageable number to screen for proteolytic activity at small scale. Also enriching for novelty seems prudent to show the potential of the metagenome compared to already published data.

4. Conclusion

From our study we highlight the possibility of mining large metagenomic datasets for proteolytic enzymes, even at subfamily level. It makes it possible to computationally screen large amounts of data and identify potential protease sequence targets for laboratory heterologous expression trials. We also included novelty enrichment to ensure that this particular sequence cannot be found in public protein databases (UniProt) or other metagenomes. With this approach the costs for expression trials can be avoided until the suitable targets are identified.

5. Methods

Samples

Arctic Ocean samples were collected in August 2009 in connection with the LOMROG II expedition at various locations. The sample locations were situated on both sides of the Lomonosov Ridge close to the pole, hence the samples represent both the Mesozoic Amerasian Basin and the Cenozoic Eurasia Basin and were obtained beneath the multi year sea ice [4]. Furthermore, one sample (sample number 20) was collected closer to Svalbard. The samples represent three areas of the water column, surface (50 – 100 m), medium (300 – 400 m) and deep (2000 – 4,300 m), which represent the epipelagic, mesopelagic and bathypelagic to abyssopelagic zone respectively. The deep samples were collected relative to the total depth which range from 2650–4460 meters and give a full representation of the water column. The samples were denoted as S (surface), M (medium) and D (deep).

Eight samples were collected at five locations in January 2007 as part of the Galathea III expedition. The five sample locations cover three different geographical areas, where samples at location 10 were located in the South Pacific, north of the ACC, the location 11 samples in the Southern Ocean, south of the ACC and location 12, 14 and 15 were sampled near the Antarctic Peninsula [23]. The sample locations near the peninsula were situated in the northern archipelago and represents a coastal environment with sample depths ranging from 400 – 1,500 m. Samples from location 10 and 11 fall under the surface, medium and deep categories mentioned above.

Detailed methods descriptions are provided in the following section. An overview of the procedure is shown in

Figure 4.

Gene finding and homology clustering

The 26 metagenomic samples were assembled with IdbauD [24] respectively. MetaGeneMark with default parameters [33] and Prodigal [11] with the universal codon table 11 were used for gene finding. Full-length genes were translated with a custom Perl script. The predicted genes of all metagenomes were pooled into one bin and homology reduced with CD-HIT-EST [17] to create the arctic marine gene catalogue. The sequence identity was set to 0.95, word size of 8, minimum length of 100 and alignment coverage for the shorter sequence of 0.9.

Gene annotation

The translated gene catalogue was aligned to Swiss-Prot release 2013_04 via BLASTp. The hits were parsed for 90% coverage over an alignment length of 50% of the query, bit score >50 and E-value <1e-05. The annotation of the best hit was resumed. The E.C. number was used for enzyme annotation and the MEROPS flag for protease annotation.

Protease-specific hidden Markov models

Hidden Markov models (HMM) models were created based on peptidases defined in MEROPS, release 9.7 [28]. The MEROPS database groups peptidases into clans, families, and subfamilies. However, to create very specific HMM models, HMMs were constructed based on individual peptidases in MEROPS (e.g. A02.063), using nearby homologues in known sequence space (UniProt release 2012.11 [2] and Novozymes internal protein database). The up to 500 nearby homologues were identified using BLASTp against the peptidases catalytic domain, using a length-dependent E-value cut-off, with a preference for including Swiss-Prot proteins with a known enzyme classification. For typical 200 – 400 amino acid domains, the E-value cutoff will be around 1e-20. A given protein could only be assigned to the closest peptidase in MEROPS, and thus never used in multiple models. The catalytic domain of each peptidase was extracted and a multiple alignment was created using MAFFT [13], and the HMMs constructed using HMMER3s hmmbuild [7]. In total, 3207 HMMs were constructed and used in the screen of the non-redundant polar marine gene catalogue.

Alignment

The HMM annotated protease sequences were aligned to UniProt release 2013.04 [2] by BLASTp and publicly available metagenomic assemblies available from the CAMERA database [30], accessed March 2013, by tBLASTn. A total of 12 metagenomic assemblies were downloaded.

Blast hits were parsed with the following parameters; coverage >50%, bit score >50 and E-value <1e-05.

Family and Subfamily diversity

The protease target sequences identified by the specific HMM approach were divided into MEROPS families and the 10 most abundant MEROPS subfamilies. The families and subfamilies were homology clustered with CD-HIT. The word size was set to 3 for sequence identities of 0.5 and 0.6, and increased to 5 for sequence identities of 0.7 to 1.0. The same was done for the 2,707 sequences after novelty enrichment.

Extracellular expression

The novelty enriched protease sequences were analyzed for evidence of extracellular expression by running SignalP 4.1 [25]. Default threshold parameters were used for parsing.

References

- [1] A. Anwar and M. Saleemuddin, "Alkaline protease from *Spilosoma obliqua*: potential applications in bioformulations.", *Biotechnology and applied biochemistry*, Vol. 31 (Pt 2), pp. 85–9, Apr. 2000.
- [2] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "The Universal Protein Resource (UniProt).", *Nucleic acids research*, Vol. 33, No. Database issue, pp. D154–9, Jan. 2005.
- [3] U. C. Banerjee, R. K. Sani, W. Azmi, and R. Soni, "Thermostable alkaline protease from *Bacillus brevis* and its characterization as a laundry detergent additive", *Process Biochemistry*, Vol. 35, No. 1-2, pp. 213–219, Oct. 1999.
- [4] J. R. Cochran, M. H. Edwards, and B. J. Coakley, "Morphology and structure of the Lomonosov Ridge, Arctic Ocean", *Geochemistry, Geophysics, Geosystems*, Vol. 7, No. 5, pp. n/a–n/a, May 2006.
- [5] D. de Pascale, C. D. Santi, J. Fu, B. Landfald, D. de Pascale, and C. De Santi, "The microbial diversity of Polar environments is a fertile ground for bioprospecting.", *Marine genomics*, Vol. 8, pp. 15–22, Dec. 2012.
- [6] M. Do Nascimento, J. C. F. Ortiz-Marquez, L. Sanchez-Rizza, M. M. Echarte, and L. Curatti, "Bioprospecting for fast growing and biomass characterization of oleaginous microalgae from South-Eastern Buenos Aires, Argentina.", *Bioresource technology*, Vol. 125, No. null, pp. 283–90, Dec. 2012.
- [7] S. R. Eddy, "Accelerated Profile HMM Searches.", *PLoS computational biology*, Vol. 7, No. 10, p. e1002195, Oct. 2011.
- [8] L. Fernández-Arrojo, "Metagenomic era for biocatalyst identification", *Current opinion in biotechnology*, 2010.

- [9] Z.-p. Guo, C.-y. Qiu, L. Zhang, Z.-y. Ding, Z.-X. Wang, and G.-Y. Shi, "Expression of aspartic protease from *Neurospora crassa* in industrial ethanol-producing yeast and its application in ethanol production.", *Enzyme and Microbial Technology*, Vol. 48, No. 2, pp. 148–54, Feb. 2011.
- [10] R. Gupta, Q. K. Beg, and P. Lorenz, "Bacterial alkaline proteases: molecular approaches and industrial applications.", *Applied microbiology and biotechnology*, Vol. 59, No. 1, pp. 15–32, June 2002.
- [11] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification.", *BMC bioinformatics*, Vol. 11, No. 1, p. 119, Jan. 2010.
- [12] B. Jaouadi, B. Abdelmalek, N. Jaouadi, and S. Bejar, "The bioengineering and industrial applications of bacterial alkaline proteases: the case of SAPB and KERAB, InTech, 2011.
- [13] K. Katoh, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic acids research*, Vol. 30, No. 14, pp. 3059–3066, July 2002.
- [14] J. Kennedy, N. D. O'Leary, G. S. Kiran, J. P. Morrissey, F. O'Gara, J. Selvin, and a. D. W. Dobson, "Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems.", *Journal of applied microbiology*, Vol. 111, No. 4, pp. 787–99, Oct. 2011.
- [15] D.-G. Lee, J. H. Jeon, M. K. Jang, N. Y. Kim, J. H. Lee, J.-H. Lee, S.-J. Kim, G.-D. Kim, and S.-H. Lee, "Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library.", *Biotechnology letters*, Vol. 29, No. 3, pp. 465–72, Mar. 2007.
- [16] Q. Li, L. Yi, P. Marek, and B. L. Iverson, "Commercial proteases: present and future.", *FEBS letters*, Vol. 587, No. 8, pp. 1155–63, Apr. 2013.
- [17] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.", *Bioinformatics (Oxford, England)*, Vol. 22, No. 13, pp. 1658–9, July 2006.
- [18] X. Liu, E. Ashforth, B. Ren, F. Song, H. Dai, M. Liu, J. Wang, Q. Xie, and L. Zhang, "Bioprospecting microbial natural product libraries from the marine environment for drug discovery.", *The Journal of antibiotics*, Vol. 63, No. 8, pp. 415–22, Aug. 2010.
- [19] M. F. Najafi, D. Deobagkar, and D. Deobagkar, "Potential application of protease isolated from *Pseudomonas aeruginosa* PD100", *Electronic Journal of Biotechnology*, Vol. 8, No. 2, pp. 79–85, 2007.
- [20] T. Nawy, "Molecular biology: Capturing sequences for bioprospecting.", *Nature methods*, Vol. 9, No. 6, p. 532, June 2012.
- [21] J. Neveu, C. Regeard, and M. S. DuBow, "Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts.", *Applied microbiology and biotechnology*, Vol. 91, No. 3, pp. 635–44, Aug. 2011.
- [22] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser.", *BMC bioinformatics*, Vol. 12, No. 1, p. 385, Jan. 2011.
- [23] A. H. Orsi, T. Whitworth, and W. D. Nowlin, "On the meridional extent and fronts of the Antarctic Circumpolar Current", *Deep Sea Research Part I: Oceanographic Research Papers*, Vol. 42, No. 5, pp. 641–673, 1995.
- [24] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data.", *Bioinformatics (Oxford, England)*, Vol. 27, No. 13, pp. i94–101, July 2011.
- [25] T. N. Petersen, S. r. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions.", *Nature methods*, Vol. 8, No. 10, pp. 785–6, Jan. 2011.
- [26] R. Prasanna, V. Gupta, C. Natarajan, and V. Chaudhary, "Bioprospecting for genes involved in the production of chitosanases and microcystin-like compounds in *Anabaena* strains", *World Journal of Microbiology and Biotechnology*, Vol. 26, No. 4, pp. 717–724, Nov. 2009.
- [27] M. B. Rao, A. M. Tanksale, M. S. Ghatge, and V. V. Deshpande, "Molecular and Biotechnological Aspects of Microbial Proteases.", *Microbiology and Molecular Biology Reviews*, Vol. 62, No. 3, pp. 597–635, Sept. 1998.
- [28] N. D. Rawlings, A. J. Barrett, and A. Bateman, "MEROPS: the database of proteolytic enzymes, their substrates and inhibitors.", *Nucleic acids research*, Vol. 40, No. Database issue, pp. D343–50, Jan. 2012.
- [29] M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman, "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms.", *Applied and environmental microbiology*, Vol. 66, No. 6, pp. 2541–7, June 2000.
- [30] S. Sun, J. Chen, W. Li, I. Altintas, A. Lin, S. Peltier, K. Stocks, E. E. Allen, M. Ellisman, J. Grethe, and J. Wooley, "Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.", *Nucleic acids research*, Vol. 39, No. Database issue, pp. D546–51, Jan. 2011.
- [31] O. P. Ward, "3.49 - Proteases", In E.-i.-C. M. Moo-Young, editor, *Comprehensive Biotechnology (Second Edition)*, pp. 571–582, Academic Press, Burlington, second edition, 2011.
- [32] T. Waschowitz, S. Rockstroh, and R. Daniel, "Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries.", *Applied and environmental microbiology*, Vol. 75, No. 8, pp. 2506–16, Apr. 2009.
- [33] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences.", *Nucleic acids research*, Vol. 38, No. 12, p. e132, July 2010.
- [34] S. B. Zotchev, O. N. Sekurova, and L. Katz, "Genome-based bioprospecting of microbes for new therapeutics.", *Current opinion in biotechnology*, Vol. 23, No. 6, pp. 941–7, Dec. 2012.

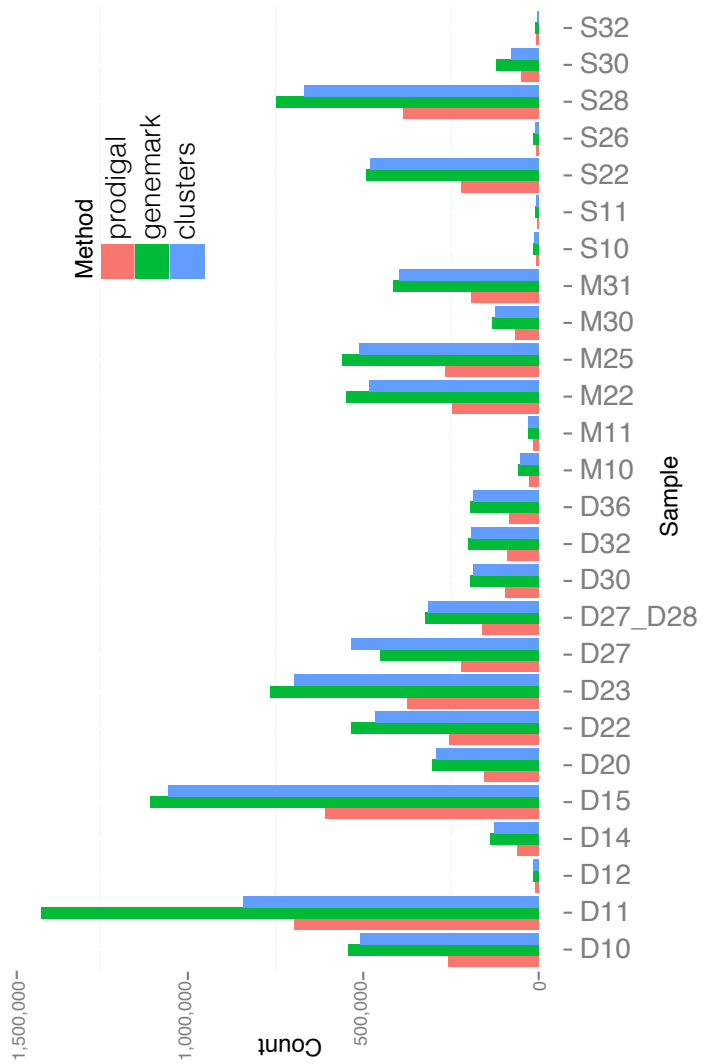


Figure 1: Genes found across samples with the gene finding programs. Prodigal and MetaGeneMark. Genes were clustered using CD-HIT

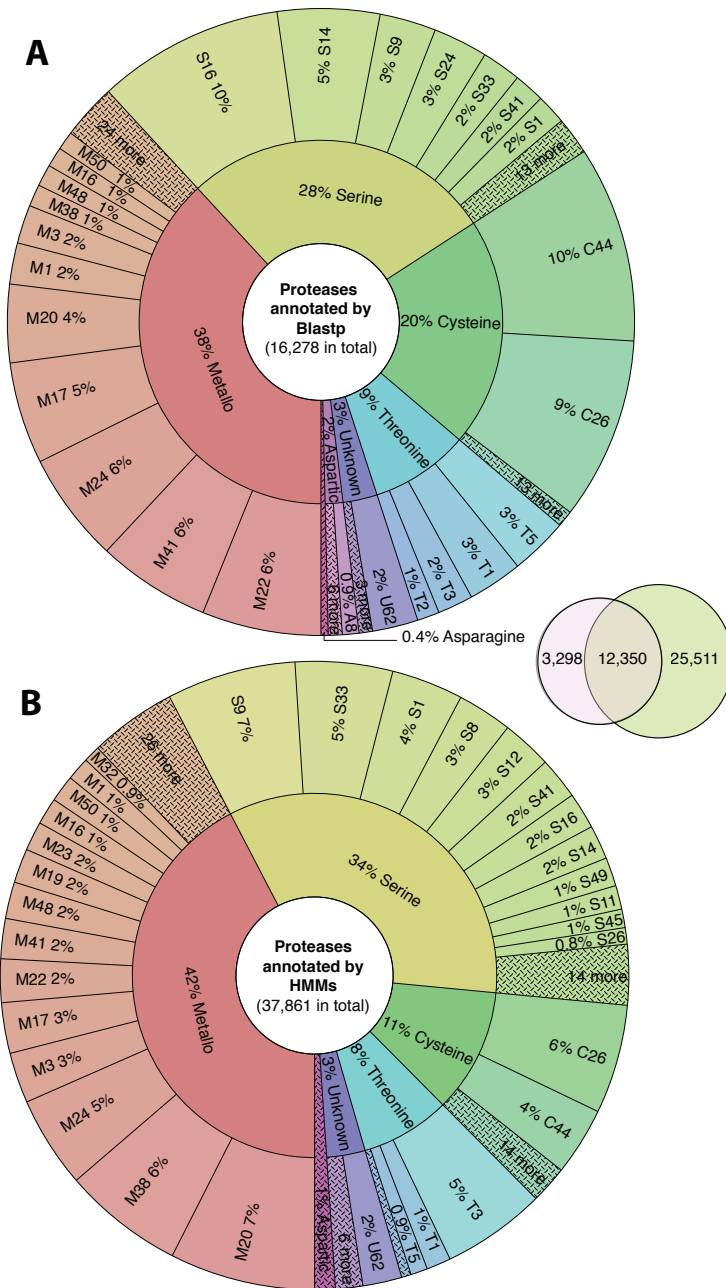


Figure 2: Swiss-Protein inferred annotation of the gene catalogue according to E.C. numbers (A) and protease HMM annotation of the translated gene catalogue (B). Sequence overlap is indicated in the middle.

Table 1: Number of aligned proteases to CAMERA metagenome assemblies and public databases

Database	Number of aligned protease sequences			
	Total	Deep	Medium	Surface
Acid Mine Drainage Metagenome	1,241	640	353	248
Waseca County Farm Soil Metagenome	9,101	4,737	1,958	2,406
Global Ocean Sampling Expedition	33,125	16,469	7,824	8,832
Mediterranean Gutless Worm Metagenome	6,349	3,366	1,379	1,604
13 Healthy Human Gut Metagenomes	9,069	4,949	1,929	2,191
Moore Marine Microbial Sequencing	25,544	13,086	5,768	6,690
Moore Marine Phage/Virus Genomes	100	46	33	21
Mouse Gut Community	432	235	79	109
Termite Gut Metagenome	1,295	703	320	272
Washington Lake Metagenomes	14,139	7,437	3,124	3,578
Whale Fall Metagenome	12,160	6,428	2,341	3,391
Swiss-Prot	15,436	7,868	3,675	3,893
TrEMBL	30,795	15,606	7,127	8,062

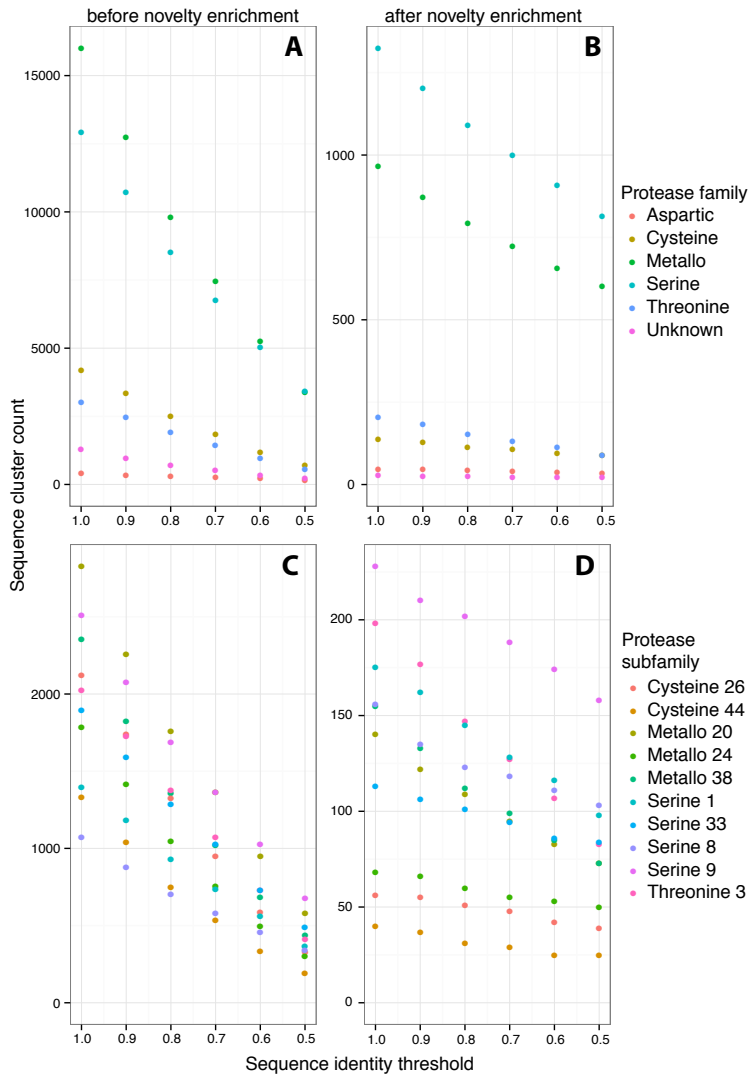


Figure 3: Sequence diversity: Sequence cluster count for sequence identity thresholds from 1 to 0.5. 37,861 HMM annotated protease sequences before and after novelty enrichment (A and B); 10 most abundant protease subfamilies before and after novelty enrichment (C and D)

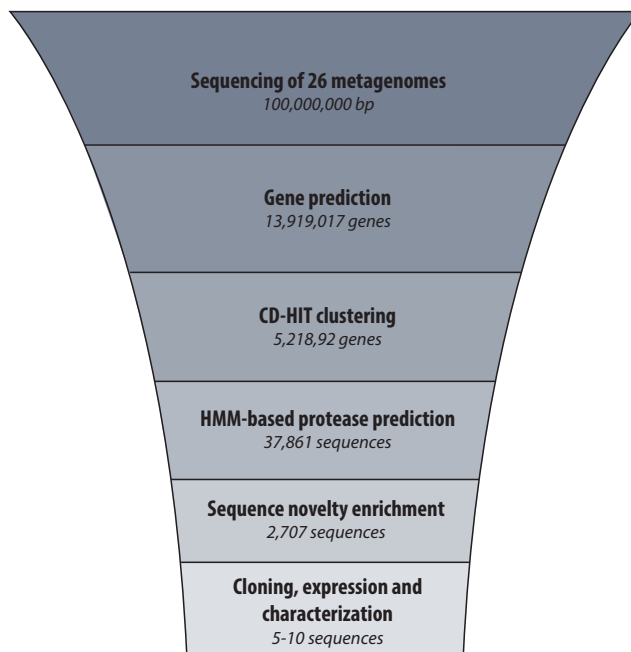


Figure 4: Schematic workflow of the major steps of protease finding in 26 polar marine metagenomes

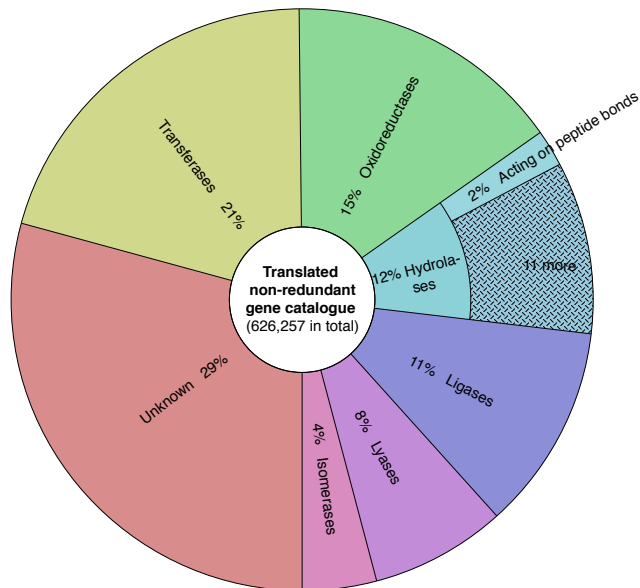


Figure S1: Polar marine gene catalogue annotation according to Swiss-Prot via BLASTp

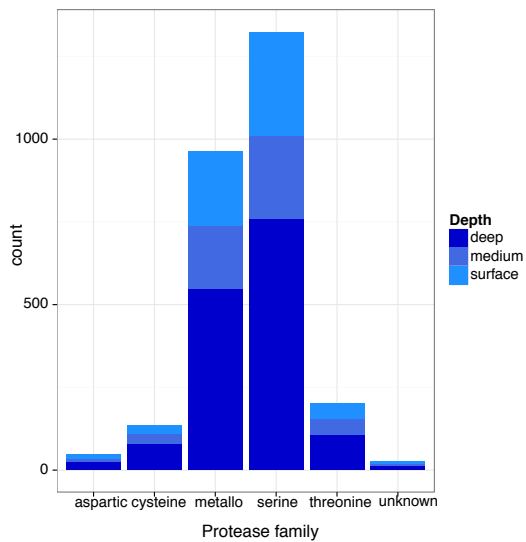


Figure S2: Number of protease sequences, which did not align to public protein databases and metagenome assemblies downloaded from CAMERA according to protease family and the depth of origin.

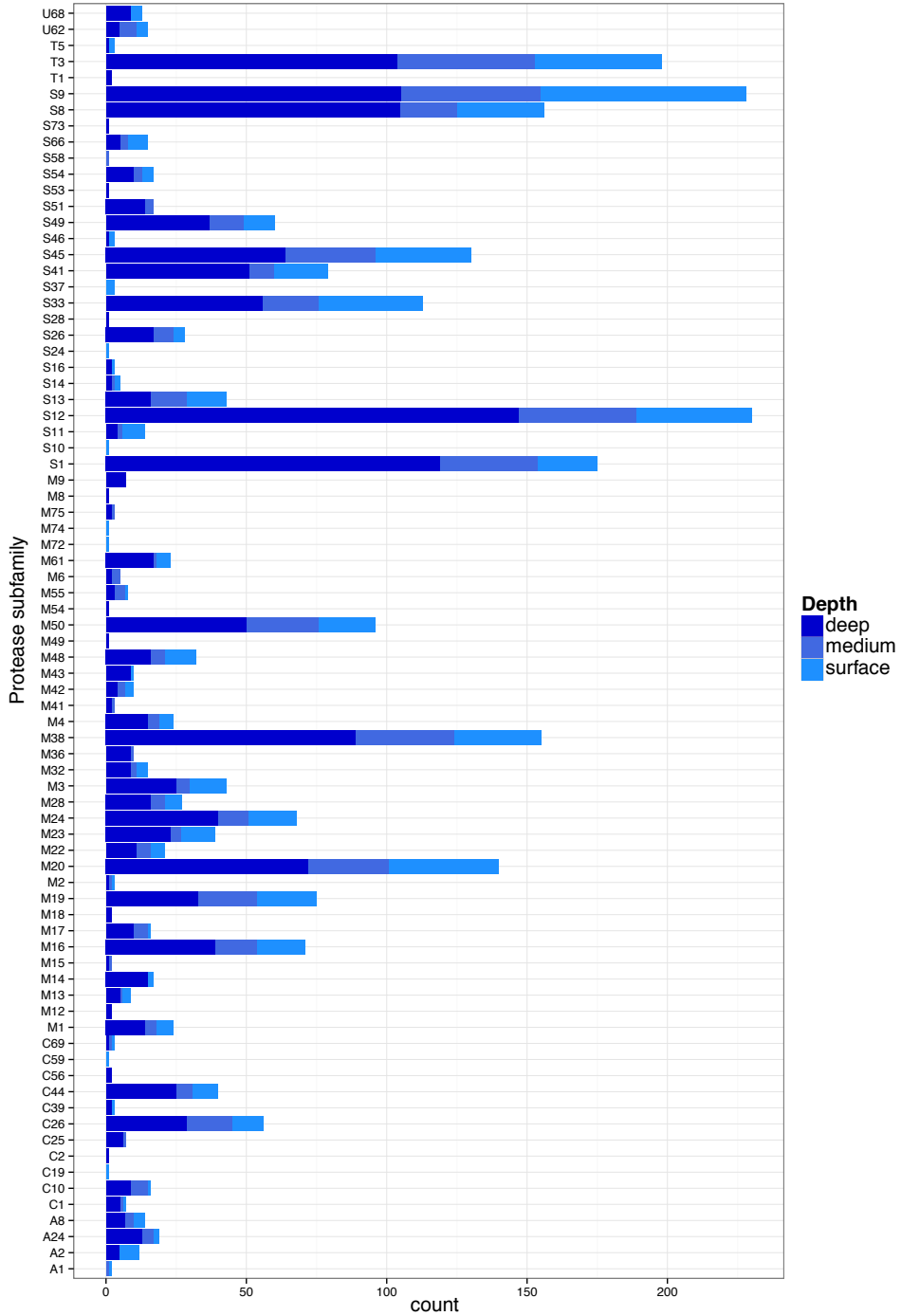


Figure S3: Number of protease sequences which did not align to public protein databases and metagenome assemblies downloaded from CAMERA according to protease subfamilies and the depth of origin

Chapter 11

Manuscript IV

11.1 Carnivorous plants - the Venus flytrap

Carnivorism within plant species is a rare trait. Only few plants are capable of digesting prey and use it as a resource for nutrients. Some 1,000 species exists which are able to trap and consume animals such as small insects and other arthropods [11]. Darwin himself expressed his fascination with the unusual adaptations of carnivorous plants by referring to it as an abominable mystery [53, 55]. His findings were published in his well-known treaties in 1875 [34].

Carnivorous plants share a common ancestor (Figure 11.1) and the carnivory trait is thought to have evolved independently six times in five different orders of flowering plants [2, 44]. Carnivorous plants have adapted to grow in places where the soil is nutrient depleted and it has been suggested to be the reason of adapting to such an uncommon lifestyle among plants [1, 21]. The most popular member of the carnivorous plant family is the renowned Venus flytrap *Dionaea muscipula*, called as 'one of the most wonderful [plants] of the world' by Charles Darwin [34]. The plant attracts insects to its brightly pigmented traps [52]. Trigger hairs need to be stimulated twice in a short succession to activate the very fast closing mechanism [52]. The plant excretes digestive fluids from glands at the inner wall into the trap to digest its prey, which are digested for up to 10 days [147]. The natural habitat of *D. muscipula* is damp pine savannas of southeastern North America, and is considered a relic species with a narrow and endangered distribution of less than 300 km² [21].

Numerous studies focused on the trap mechanism of the venus flytrap which is unique due to its motion sensor mechanism [15, 52, 91, 181]. However, no genome or mixed tissue transcriptome data have been available.

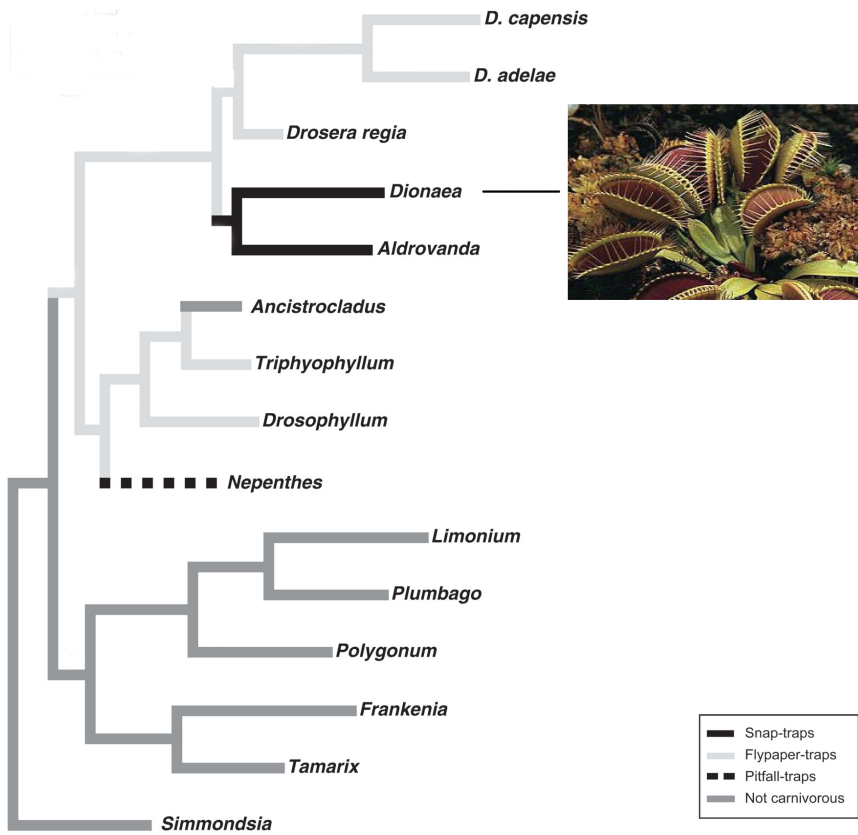


Figure 11.1. Phylogenetic relationships for the carnivorous plant genera in Caryophyllales inferred from parsimony analysis; Branch coloring represents trait: not carnivorous (dark gray), pitfall-traps (dashes), flypaper-traps (light gray), and snap-traps (solid black). Modified from Cameron *et al.* [21].

Schulze *et al.* [146] used transcriptome data to delineate the protein composition of the digestive fluid of *D. muscipula*. Their study showed that the digestive fluid system is similar to other carnivorous plants such as *Nepenthes* (a tropical pitcher plant [113]), however, the proteolytically active enzymes of *Nepenthes* and also vertebrates are predominantly aspartic proteases [7, 80, 146]. In the digestive fluid of the Venus flytrap, cysteine proteases are the most abundant class of proteases, followed by a serine carboxypeptidase and aspartic proteases [146].

In the present study, we sequenced the transcriptome of *D. muscipula*, using a mixed-tissue sample for cost-effective deep sequencing of a normalized cDNA library, complementing the selection of accessible transcriptome data. The transcriptome sequences were assembled into contigs. Functional annotation and gene ontology analyses were performed, and a large number of transcripts related to catalytic activities were identified. This is the first high throughput data publicly available for a member of the largest family of carnivorous plants (Droseraceae). Our data provide a public resource for unveiling mechanistic features of the carnivorous syndrome such as attraction, trapping and digestion. Moreover, to expand the list of genome size estimates of members of the carnivorous orders, we present the first genome size estimate of a member of the sundew family in the order Caryophyllales. The supplementary sequences can be found in the Appendix.

11.2 Transcriptome and genome analyses of the Venus flytrap (*D. muscipula*)

Transcriptome and genome analyses of the Venus flytrap (*Dionaea muscipula*)

Michael Krogh Jensen^{1*}, Josef Korbinian Vogt^{2*},
Simon Bressendorff¹, Andaine Seguin-Orlando³, Hamed El-Serehy⁴,
Morten Petersen¹, Khaled Al-Rasheid⁴, Thomas Sicheritz-Pontén², John Mundy¹

¹Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark

²Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark

³National High-throughput DNA Sequencing Centre, ster Farimagsgade 2D, DK-1353 Copenhagen, Denmark

⁴Zoology Department, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

* Joint first authorship.

ABSTRACT – Background: The Venus flytrap (*Dionaea muscipula*) is renowned from Darwin's early studies on plant carnivory and the origins of species. A fascinating feature of *D. muscipula* is its rapid snap trapping movement triggered by mechanical stimulation of specialized sensory hairs. To provide tools to further analyze the evolution and functional genomics of *D. muscipula*, we sequenced a normalized cDNA library from mRNAs of snap traps and flowers, and assembled a basal transcriptome. As earlier studies identified great variation in genome size among members of a single carnivorous family, we also determined the genome size of *D. muscipula*.

Results: We sequenced a normalized cDNA library synthesized from mRNA isolated from *D. muscipula* flowers and traps. Using the Oases transcriptome assembler we assembled 79,165,657 quality trimmed reads into 80,806 cDNA contigs, with an average length of 679 bp and an N50 length of 1,051 bp. A total of 17,047 unique proteins were identified, and assigned to Gene Ontology (GO) and classified into functional categories. A total of 15,547 full-length cDNA sequences were identified, from which open reading frames were detected in 10,941. Comparative GO analyses revealed that *D. muscipula* is highly represented in molecular functions related to catalytic, antioxidant, and electron carrier activities. Also, using a single copy sequence PCR-based method we estimate that the genome size of *D. muscipula* is ~3 Gb, almost 50 times larger than that of carnivorous *Genlisea margaretae*.

Conclusion: We present the sequencing, assembly and functional annotation of a normalized transcriptome of *D. muscipula*. We highlight the quality of normalized cDNA libraries to cost-effectively provide good coverage of both low and high abundant transcripts of Gb-sized genomes such as *D. muscipula*. Our genome and transcriptome analyses will contribute to future research on this fascinating, monotypic species and its heterotrophic adaptations.

KEY WORDS – Venus flytrap, transcriptome, annotation, genome size

1. Introduction

Darwin was fascinated by the unusual adaptations of carnivorous plants during his often frustrating studies of the evolution of flowering plants which he referred to as an abominable mystery [13, 14]. Darwin published his treatise on insectivorous plants after roughly a decade of study [10]. In this work he noted that the Venus flytrap (*Dionaea muscipula*) was one of the most wonderful of the world. Studies of carnivorous plants have continued since Darwin's time. Attention has focused on the biogeography and

phylogenetics of the only two carnivorous species with snap traps, *D. muscipula* and the aquatic waterwheel *Al-drovanda vesiculosa* [5, 9, 23, 25]. The natural habitat of *D. muscipula* is damp pine savannas of southeastern North America, and is considered a relic species with a narrow and endangered distribution of less than 300 km² [9]. *A. vesiculosa* is also considered a relict earlier widely distributed in Europe, Africa, India, Japan, and Australia, yet now confined to fewer than 36 localities mostly in Europe and Russia [1].

Earlier molecular phylogenetic studies demonstrated that

carnivory occurs in several flowering plant lineages [3, 19], and it was thought that the snap traps of *A. vesiculosa* and *D. muscipula* may have evolved independently. However, their unique snap traps are not examples of convergent evolution, but share a common, old-world ancestor at least 65 million years ago [9, 21, 25]. More precisely, Cameron *et al.* [9] used sequences from nuclear 18S and plastid *rbcL*, *atpB*, and *matK* genes to show that *A. vesiculosa* and *D. muscipula* evolved as monotypic sister genera from a sundew-like ancestor. While the habitat of *A. vesiculosa* is similar to that of many aquatic carnivorous bladderworts (*Utricularia* spp.), the snap traps of *D. muscipula* and *A. vesiculosa* are unique in having a single evolutionary origin, and narrow ecological distributions [14].

Improved understanding of the molecular adaptations to plant carnivory has also been sought via genome size estimates. Interestingly, genome size varies more than 2,300-fold among angiosperms, from that of *Paris japonica* ($2n = 12$, $1C = 152.20$ pg DNA or ~ 149 Gbp [22]), to that of carnivorous *Genlisea margaretae* ($2n = \sim 40$, $1C = 0.0648$ pg or ~ 63 Mbp [15]). The biological significance of this massive variation is puzzling. Carnivorous plants are found in at least five, genetically poorly described orders [12]. The lack of molecular tools and genetic information, however, has not hampered phenotypic and ecological studies of the orders with carnivorous members [2, 14], and comparative genomic analyses can clarify a number of their traits. Within the Lentibulariaceae, Greilhuber *et al.* [15] identified ~ 24 -fold variation in genome sizes among *Genlisea* and other family members. Also, large variations in ploidy levels and chromosome sizes have been reported within the carnivorous Droseraceae [16], and Rogers *et al.* recently reported genome estimates for two carnivorous pitcher plants, *Sarracenia purpurea* and *Sarracenia psitticina*, to be larger than 3.5 Gb [26]. Thus, carnivorous plants seem to have an extreme plasticity in terms of genome content, and such large genomes tend to have many repetitive sequences and transposable elements [26].

An important complement to genome size analyses comes from transcriptome data. Both transcriptome and genome sequence data are needed to understand the physiological and genetic basis of the snap trap and to identify genes selected during its evolution [24]. To this end, deep sequencing is beginning to reveal certain aspects of the evolution of carnivory. Recently, transcriptome data for the bladderwort *Utricularia gibba* was published using next-generation sequencing [17], and Srirastava *et al.* [31] have reported the deep sequencing of two *Sarracenia* species, providing valuable information on the events of genome duplication and speciation within the genus *Sarracenia*. Finally, Schulze *et al.* [29] used transcriptome data to delineate the protein composition of the digestive fluid of *D. muscipula*. Such studies pave the way to understand the molecular physiology associated with features of the carnivorous syndrome.

Table 1. Statistics of transcriptome sequencing and assembly of *D. muscipula*

Sequencing	# of reads (93 bp single-end)	81,329,943
	Total bases	7.56 Gb
	# cleaned reads	79,165,657
Assembly	# of contigs	80,806
	Max contig length	7,545 bp
	Min contig length	100 bp
	Mean contig length	679 bp
	N50 length	1051 bp

In the present study, we sequenced the transcriptome of *D. muscipula*, using a mixed-tissue sample for a cost-effective deep sequencing of a normalized cDNA library. The transcriptome sequences were assembled into contigs. Functional annotation and gene ontology analyses were performed, and a large number of transcripts related to catalytic activities were identified. This is the first high throughput data publicly available for a member of the largest family of carnivorous plants (Droseraceae), the renowned *D. muscipula*. Our data provide a public resource for unveiling mechanistic features of the carnivorous syndrome such as attraction, trapping and digestion. Moreover, to expand the list of genome size estimates of members of the carnivorous orders, we present the first genome size estimate of a member of the sundew family in the order Caryophyllales.

2. Results

2.1 Transcriptome Sequencing and Assembly of *D. muscipula*

To analyse the transcriptome of *D. muscipula*, a normalized library of mixed mRNAs from traps and flowers was sequenced using Solexa HiSeq2000 sequencing technology. A total of 81,329,943 single-end reads were generated with a read length of 93 bp (excluding Illumina barcode index). After removal of ambiguous nucleotides and low-quality sequences (Phred quality score < 20), a total of 79,165,657 cleaned reads (97.3%) were obtained. These raw transcriptome sequences in this study have been deposited in the NCBI SRA database (Accession number SRA091387), and recovered reads were assembled. As shown in Table 1, the transcriptome was assembled, combining 79,165,657 reads into 80,806 contigs, ranging from 100 to 7,545 bp in length. The average length was 679 bp, and the N50 length was 1,051 bp.

To quality assess contig assemblies and validate our normalization procedure, we selected 10 contigs for PCR-based validation. The contig were selected based on the

alignment annotation to putative low- and high-abundant transcript genes. For high-abundant mRNA transcripts this included actin and ubiquitin sequences, and for putative low-abundant mRNA transcripts it included transcription factor sequences. Primers were designed to target a range of contig sizes, and to span a range of putative mRNA abundances. Using an independent biological replicate cDNA template of *D. muscipula* traps and flowers, we validated transcript assemblies ranging from 247-1014 bp (Figure 1, and Supplemental file S1), including both putative low- and high-abundant transcripts. Expected amplicon sizes were obtained from all ten contigs, although no genomic amplicon was obtained for *DmUCH-like* (Supplementary file S1). In conclusion, this confirmed that the assembly using the Oases algorithm was reliable, and that our normalization procedure enabled identification of transcript abundances with an apparently large dynamic range.

2.2 Functional Annotation

The assembled contigs were aligned to the NCBI non-redundant (nr) protein database for functional annotation by BLASTx with an E-value cut-off of $1e-5$. A total of 42,656 contigs had a significant hit, corresponding to 17,047 unique protein accessions in the nr protein database (Table 2).

Gene ontology (GO) analysis was conducted on these 17,047 unique proteins using InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>) on integrated protein databases with default parameters. A total of 9,909 unique proteins were assigned to at least one GO term for describing biological processes, molecular functions and cellular components. The InterProScan output file was input to the BGI WEGO program and GO annotations plotted (<http://wego.genomics.org.cn>) (Figure 2). Briefly, in the cellular component division, genes related to cell parts and macromolecular complexes (2,588 (26.3%) GO:0044464 and 746 (7.6%), GO: 0032991, respectively) are highly represented. Interestingly, in contrast to other plants, *D. muscipula* also has genes related to a virion part (3 (0.1%), GO:0044423). For the molecular function division, a large abundance of genes are related to binding and catalytic activity (5348 (54.4%) GO:0005488 and 4847 (49.3%) GO:0003824, respectively). Also, antioxidant (56 (0.6%) GO:0016209) and electron carrier activities (184 (1.9%) GO:0009055) are represented. For the biological process division, genes involved in cellular (4,285 (43.6%), GO:0009987) and metabolic processes (5,136 (52.2%), GO:0008152) are highly represented, including the child term of establishment of localization (733 (7.4%), GO:0051234). In contrast, genes associated with developmental and multicellular organismal processes were lowly represented (6 (0.1%), GO:0032502; and 14 (0.1%) GO:0032501, respectively), compared to full-genome annotations for *Arabidopsis* (15%

and 15.5%, respectively). This may well reflect the limited tissues and developmental stages sampled in this study. The complete GO annotation results are shown in Supplementary file S2.

2.3 Assessment of Transcriptome Assembly

The assembled transcript contigs were aligned to all RefSeq entries for a moss (*Physcomitrella patens*), and the angiosperms grape (*Vitis vinifera*), *Arabidopsis thaliana*, tomato (*Solanum lycopersicum*), *Brachypodium distachyon*, rice (*Oryza sativa*), maize (*Zea mays*), and the monotypic oil plant *Ricinus communis* using BLASTx with an E-value cutoff of $1e-5$ (Table 2). Cross-species sequence similarity showed most hits identified in grapes, tomatoes, oil plants and *Arabidopsis*. When looking into unique protein hits, the *D. muscipula* transcriptome, originating from a normalized mixed-tissue cDNA library, targeted almost 60% of the tomato RefSeq entries, and more than 50% of the entire grape RefSeq data. Likewise, almost 50% of the *Brachypodium* RefSeq data was uniquely aligned by individual *D. muscipula* contigs. For *Arabidopsis*, 13,469 unique protein hits were identified, covering more than one third of the total number of *Arabidopsis* RefSeq protein entries. These numbers represent underestimates of the minimal number of *D. muscipula* genes found expressed in the two tissues used in this study, flowers and traps. Apart from tissue-specificity, it is possible that many *D. muscipula* unique protein hits could not be aligned to RefSeq hits because they represent untranslated regions (UTRs) and/or non-coding RNAs (ncRNAs). All together, the high numbers of unique protein hits aligned by *D. muscipula* contigs underscores the quality of the data obtained from our mixed-tissue and normalized library.

2.4 Full-Length cDNA prediction

Full-length cDNAs are important resources for many applications, including reverse genetic and evolutionary studies. To search for potentially full-length cDNAs with complete open-reading frames (ORFs) in the assembled *D. muscipula* transcriptome, all contigs were analyzed by TargetIdentifier [20]. A total of 15,547 full-length sequences were identified from the assembly. The size distribution of full-length sequences compared to the size distribution of our total 80,806 cDNA contigs is presented in Figure 3. In contrast to the size distribution of the total contig number, full-length sequences are biased towards those > 1 kb in length. This indicates that short full-length cDNA sequences may be underrepresented in our assembly and transcriptome data.

2.5 Genome Size Estimates

An intriguing observation from genome studies of carnivorous plants is the extreme size differences observed even among individual family members [15]. To expand the list of genome size estimates of members of the carnivorous orders, we estimated the genome size of *D. muscipula*. Using an improved protocol adapted from Bekesiova *et al.* [6], we obtained high-quality genomic DNA. Figure 4A shows an example of 200 ng DNA extracted using this method. We routinely obtained approx. 25 and 50 µg high quality (A260/280 > 1.8, and A230/260 > 2.0) genomic DNA per g fresh weight from traps and flowers, respectively.

To estimate the genome size of *D. muscipula* using the qPCR-based method of Wilhelm *et al.* [34], a DNA sample without significant RNA contamination is required. From purified gDNA, we targeted the amplification of a single-copy genic region assembled and validated (see Figure 1) from our *D. muscipula* transcript sequencing. With this sequence as a query we used BLASTx to identify the closest homologue. This identified *Arabidopsis ACTIN7 (ACT7)* as the closest homolog, with total query coverage of 67% and maximum shared identity of 86% (Supplementary file S3). We therefore designated this target *D. muscipula* amplicon *DmACT7*. Using this amplicon, the genome size for *D. muscipula* was estimated to be 2956 Mbp (SEM= 210 Mbp, n=11), equivalent to 3.02 +/- 0.21 pg for the 1C haploid genome (Figure 4B, Table 3). As a control, we estimated the genome size of the model angiosperm *A. thaliana* using the *ACTIN1 (ACT1)* genic region. This estimate of 173 Mbp (SEM= 21 Mbp, n=7; Figure 4B and Table 3) overlaps the well-documented value of the *A. thaliana* genome of 157 Mbp (0.16 pg; [4, 7]).

3. Discussion

To date, the highest diversification rates among angiosperms are found in the order Lamiales [37]. In particular, the apparent plasticity observed in the large Lentibulariaceae family has recently received attention [2, 15]. In this carnivorous family, three taxa exhibit significantly lower IC-values than the 157 Mbp of *Arabidopsis thaliana*. These are *Genlisea margaretae* with 63 Mbp, *G. aurea* with 64 Mbp, and *Utricularia gibba* with 88 Mbp [15]. Our size estimate for the Droseraceae family member *D. muscipula* is 46-fold higher than that of the *G. margaretae* genome, and comparable to the genome size estimates for carnivorous pitcher plants [26]. Such estimates enable calculation of the minimum number of high-quality reads required for whole-genome sequencing of *D. muscipula* and other Gb-sized genomes from carnivorous plants. A good sequencing coverage should provide reliable information on the evolution of carnivory.

The estimated haploid genomes of 3-4 Gbp indicate that certain carnivorous plants have undergone dramatic genome evolution. An explanation for such massive proliferation of genome rearrangements, as observed in plastid genomes of Lentibulariaceae members, may be associated with increasingly relaxed functional constraints due to the heterotrophic lifestyle of carnivorous plants [21, 30, 33]. Another explanation is that high nucleotide substitution rates are linked to reactive oxygen species (ROS) generated from the increased respiratory rates needed for the oxidative phosphorylation of ADP to ATP upon movement of trapping devices in carnivorous plants [2, 18]. ROS is known to cause oxidation of nucleotide bases and generation of DNA strand breaks [8].

With respect to the *D. muscipula* transcriptome, *D. muscipula* shares the greatest sequence similarity to tomato (59.8%, Table 2). This is not a surprise, as tomato is the only species included from the asterids clade, to which *D. muscipula* also belongs. However, the assembled transcriptome of *D. muscipula* also shares large sequence similarities to the rosids clade member *Vitis vinifera* (53.8%, Table 2). The relatively strong sequence similarity between carnivorous species and grapes was also reported in a transcriptome study of the carnivorous pitcher plants, *Sarracenia psittacina* and *Sarracenia purpurea* [31]. Future sequencing data on more asterids and rosids members, including transcriptome comparisons with other carnivorous species [17, 29, 30] will aid the research community to delineate the intriguing phylogeny and molecular adaptation of carnivorous plants and their ecology.

We note that our cost-effective approach using a normalized library of mixed tissues from trap and flowers was only collected from adult plants. Our data therefore does not cover the whole *D. muscipula* transcriptome. Still, our data aligned almost 50-60% of the entire complement of RefSeq entries for several model and crop species. Future studies may address the identification of tissue and developmentally regulated genes by temporal and spatial sampling of tissues under different conditions. At present, our data may be mined for comparative studies and as an annotative tool for whole-genome sequencing and future *de novo* assembly of the *D. muscipula* genome.

4. Conclusion

In this study, the transcriptome of *D. muscipula* was sequenced, *de novo* assembled and functionally annotated. An ORF analysis was conducted and a large number of full-length cDNA sequences were identified. The *D. muscipula* transcriptome provides some insight into the molecular processes occurring in a Gb-sized carnivorous plant genome. Abundant representation of processes related to the expression of genes associated with binding, catalytic, antioxidant and electron carrier activities was observed.

Future uniform meta-analyses of short-read archives, including cDNA sequences from carnivorous *Utricularia* [17] and *Sarracenia* [31] species will aid future studies of carnivorous plants and their ecology. This underlines the importance of further expansion of sequence repositories, especially for non-model organisms, for improved understanding of molecular physiology and evolution related to Darwins abominable mystery.

5. Methods

Plant material

For nuclear genome estimates, 1 g of freshly harvested flowers, petioles and traps were used from *D. muscipula* and *Arabidopsis thaliana* (Col-0). *D. muscipula* plantlets were purchased from Horticulture Lammehave A/S (Ringe, Denmark).

Genomic DNA extraction

DNA was extracted from *D. muscipula* and *A. thaliana* as described for *Drosera rotundifolia* by Bekesiova *et al.* [6] with modifications for extraction from the more succulent and recalcitrant *D. muscipula*. After tissue grinding, plant cells were lysed in 6 ml CTAB-buffered N-lauryl sarcosine (5%) with 2 μ l 2-mercaptoethanol and 0.3 g polyvinylpyrrolidone (PVPP), (MW=360,000, Sigma) per ml lysis buffer, and incubated 1 hr at 65°C in a water bath. PVPP and 2-mercaptoethanol respectively bind to and remove polyphenols, polysaccharides and tannins from plant extracts. Following lysis, the lysate became more viscous as the solution was cooled at room temperature for 10 min before extraction with 1 x volume of chloroform:isoamyl alcohol (24:1). The sample was then centrifuged at 13,000 RPM for 10 min at 4°C. A 5-ml pipette was used to gently transfer the upper aqueous phase to new tubes and the DNA was precipitated overnight at -20°C using 0.1 volumes of 3 M Na-acetate (pH 5.2) and 2.5 volumes ice-cold ethanol. The precipitated DNA was collected by centrifugation for 20 min at 13,000 RPM and 4°C. The pellet was washed in 70% ethanol and centrifugation repeated. The pellet was briefly air-dried at room temperature before being gently dissolved in 1 ml TE (pH 7.5). At this point, due to high absorbance at 230 nm, a second purification round was used. First, resuspended DNA was treated for 1 hr at 37°C with 50 μ g/ml RNase A (Sigma) and 50 units/ml RNase T1 (Fermentas). Proteinase K (150 μ g/ml) was then added for another hour at 37°C. Subsequently, 1 x volume of CTAB buffer was added and the solution incubated 1 hr at 65°C. 1 ml of chloroform:isoamyl alcohol (24:1) was then added and mixed. After centrifugation for 10 min at 13,000 RPM at 4°C, the supernatant was transferred to a new tube and again precipitated over-night at -20°C with 0.1 volumes of

3 M Na-acetate (pH 5.2) and 2.5 volumes ice-cold ethanol. DNA was then collected for 20 min at 13,000 RPM and 4°C. The pellet was washed in 70% ethanol and centrifugation repeated. The pellet was air-dried for 30 min at room-temperature and resuspended in TE (pH 7.5) or water. DNA purity and concentration was measured on a nanodrop 1000 (Thermo scientific). For pure DNA, the A260/280 ratio should be > 1.8.

mRNA isolation

Total RNA was extracted from 1.5 g fresh weight each of *D. muscipula* flowers and traps using an optimized urea-based protocol. For a single extraction, 0.1 g (approx. equivalent to 1 medium sized trap) plant material was flash-frozen in liquid nitrogen and ground together with 0.03 g of PVPP. Plant powder was then transferred to a pre-warmed (65°C) microcentrifuge tube containing 700 μ l of RNA extraction buffer (2% CTAB (w/v), 2% PVP K25 (w/v), 100 mM Tris-HCl (pH 8.0), 25 mM sodium-EDTA (pH 8.0), 2.0 M NaCl, 2% (w/v) β -mercaptoethanol (add just before use)) and vigorously shaken. The suspension was then centrifuged for 2 min at 13,000 RPM to pellet plant debris, and the supernatant transferred to a new tube. Subsequent steps are all performed at 4°C. Total RNA was then extracted with 600 μ l of chloroform:IAA (24:1), and the phases separated by centrifugation (10,000 RPM, 10 min., 4°C). The top aqueous phase was transferred to a new microcentrifuge tube and extracted with 500 μ l of phenol:chloroform:IAA (25:24:1) using centrifugation (10,000 RPM, 10 min., 4°C). Transfer top aqueous phase to a new tube and 0.25 volumes added (125 μ l to 500 μ l) of 10 M LiCl and gently mixed well. RNA was precipitated overnight at 4°C, and then pelleted by centrifugation (10,000 RPM, 20 min, 4°C). Dissolve RNA in 100 μ l of DEPC-treated water. Samples are then re-precipitated by 250 μ l precipitation mix (80% EtOH, 20% 1M sodium acetate (pH 5.2)) and incubated 1 hr at -70°C. Subsequently, RNA is centrifuged (10,000 RPM, 20 min., 4°C). The pellet was gently washed in 70% RNase-free EtOH, centrifuged (10,000 RPM, 20 min., 4°C) and resuspended in 30 μ l DEPC-treated water. Subsequently, total RNA was RQ1 DNase treated (Promega), and mRNA isolated from 2-3 mg of trap and flower total RNA using PolyATtract (Promega), according to the manufacturers description.

cDNA library construction, sequencing and assembly

The MINT kit (Evrogen) was used for first-strand cDNA synthesis with 400 ng mRNA from each sample. Following evaluative PCR, a full-sized pre-saturation synthesis of ds-cDNA was prepared for both tissues using Encyclo PCR (Evrogen). cDNA was purified using QIAquick (Qiagen) and concentration measured using Qubit (Invitrogen). Samples were then pooled in a 1:4 ratio of

trap:flower cDNA to a total of 1 µg cDNA for normalization using duplex-specific nuclease [35, 36]. Normalization was evaluated by PCR using Evrogen PCR adaptor-specific primer M1, and a full-size cDNA amplification performed. A total of 4 µg cDNA was subsequently fragmented using a Bioruptor® (Diagenode) and MinElute (Qiagen) purified prior to library building. The NEBNext® Quick DNA library kit (New England Biolabs) was used for library building with 0.5 µg fragmented cDNA and 1 µl of 15 µM InPE adaptor (Illumina). Following another MinElute step we indexed (6-bases) and amplified the library 10x with Illumina standard primers (InPE1.0 and InPE2.0). Finally, the library was evaluated by gel electrophoresis and a gel piece containing 270-320 bp fragments was isolated and QIAquick purified (Qiagen). The library was sequenced using Solexa HiSeq2000 sequencing technology with 101 bp single-end reads at the National High-throughput DNA Sequencing Centre, University of Copenhagen. All sequenced reads have been uploaded to NCBI Short Read Archive with accession number SRA091387. Prior to *de novo* assembly using Oases [28], adaptor sequences were trimmed and low quality reads removed (Phred quality score < 20) by in-house software including the FASTX-Toolkit available from http://hannonlab.cshl.edu/fastx_toolkit/. To quality assess the transcriptome assembly, the contigs were aligned by BLASTn (E-value ≤ 1e-5) and 10 contigs with varying size were selected for assembly confirmation. Primers were designed using Primer3 [27]. Sequences for primers and contigs can be found in Supplementary file S1.

Functional annotation

The assembled transcriptome contigs were aligned to NCBI non-redundant protein databases (nr) using BLASTx (E-value ≤ 1e-05, bit score ≥ 50). Gene names and annotation were assigned to the corresponding contig based on the best BLASTx hit. Transcripts for each locus were scanned with InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>) on integrated protein databases with default parameters. The GO terms associated to the transcriptome contigs were retrieved to describe genes in the categories of cellular components, molecular function and biological process. The functional gene annotation for *Arabidopsis* was retrieved from The Arabidopsis Information Resource, version TAIR10 [32].

Assembly assessment and full-length cDNA identification

The assembled contigs were aligned to non-redundant protein databases with BLASTx at a cut-off E-value of 1e-5 and putative full-length cDNA sequences and ORFs were identified by TargetIdentifier (<http://proteomics.yzu.edu/tools/TargetIdentifier.html>) [20]. cDNA sequences are classified as full-length if the following criteria were fulfilled (1) the sequence has a start codon with a downstream stop

codon or (2) the sequence has a stop codon and an in-frame start codon is detected prior to the 10th codon of the aligned subject sequence. For comparison of the *D. muscipula* open reading frames to other plant proteins, the contigs were aligned with BLASTx (standard parameters, E-value < 1e-5) to 8 RefSeq and Ensembl proteins, including *Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa*, *Physcomitrella patens*, *Ricinus communis*, *Vitis vinifera*, *Solanum lycopersicum* and *Zea mays*. Hits were parsed with standard parameters and the best hit was resumed.

qPCR estimate of genome size

Sequencing of the transcriptome of traps and flowers of *D. muscipula* gave a total of 80,806 contigs. A long unique sequence with good coverage was chosen for primer design as shown in Supplementary file S3. The sequence had 86% identity to the *Arabidopsis ACT7* gene (AT5G09810). Primers were from MWG Biotech (Ebersberg). The sequence and positions of primers are shown in Supplementary file S3. The qPCR-based analysis of genome size was performed according to Wilhelm *et al.* [34] using a Bio-Rad iCycler (Bio-Rad). The genome size, described as gametic nuclear DNA contents (C-values), either in units of mass (picograms, where 1 pg = 10⁻¹² g) or in number of base pairs (where 1 pg DNA = 0.978 × 10⁹ bp; [11]), was calculated by dividing the mass of sample DNA by the copy number determined for single copy genes.

Acknowledgements

JM, KAR, HE-S and MKJ were supported by the Distinguished Scientist Program of King Saud University, Riyadh, Saudi Arabia. No individuals employed or contracted by the funders, other than the named authors, played any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting Information

Supplementary file S1 (Sequences used for assembly validation) Supplementary file S2 (Complete GO annotation term summary) Supplementary file S3 (DmACT7 sequence, including primer locations)

Author Contribution

Conceived and designed the experiments: MKJ, SB, JM, MP, and TSP. Performed the experiments: MKJ, SB, AS-O, and JKV. Analyzed the data: MKJ, SB, JM, and JKV. Contributed reagents/materials/analysis tools: JM, MP, HE-S, and KAR. Wrote the paper: MKJ, JM, TSP, and JKV.

References

- [1] L. Adamec, "Ecological requirements and recent European distribution of the aquatic carnivorous plant *Aldrovanda vesiculosa* L.A review", *Folia Geobotanica*, 1995.
- [2] V. Albert and R. Jobson, "The carnivorous bladderwort (*Utricularia*, *Lentibulariaceae*): a system inflates", *Journal of Experimental Botany*, 2010.
- [3] V. A. Albert, S. E. Williams, and M. W. Chase, "Carnivorous plants: phylogeny and structural evolution.", *Science*, Vol. 257, No. 5076, pp. 1491–5, Sept. 1992.
- [4] G. Arabidopsis, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.", *Nature*, 2000.
- [5] T. Bailey and S. McPherson, *Dionaea: The Venus's Flytrap*, Redfern Natural History Productions Ltd, 2012.
- [6] I. Bekesiova, J. Nap, and L. Mlynarova, "Isolation of high quality DNA and RNA from leaves of the carnivorous plant *Drosera rotundifolia*", *Plant Molecular Biology Reporter*, 1999.
- [7] M. D. Bennet, "Comparisons with *Caenorhabditis* (100 Mb) and *Drosophila* (175 Mb) Using Flow Cytometry Show Genome Size in *Arabidopsis* to be 157 Mb and thus 25 % Larger than the *Arabidopsis* Genome Initiative Estimate of 125 Mb", *Annals of Botany*, Vol. 91, No. 5, pp. 547–557, Feb. 2003.
- [8] A. Britt, "DNA damage and repair in plants", *Annual review of plant biology*, 1996.
- [9] K. Cameron, "Molecular evidence for the common origin of snap-traps among carnivorous plants", *American Journal of Botany*, 2002.
- [10] C. Darwin, *Insectivorous Plants*, John Murray, London, 1875.
- [11] J. Dolezel, J. Bartos, H. Voglmayr, and J. Greilhuber, "Nuclear DNA content and genome size of trout and human.", *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, Vol. 51, No. 2, pp. 127–8; author reply 129, Feb. 2003.
- [12] A. Ellison and N. Gotelli, "Energetics and the evolution of carnivorous plants Darwin's 'most wonderful plants in the world'", *Journal of Experimental Botany*, 2009.
- [13] W. Friedman, "The meaning of Darwin's abominable mystery", *American Journal of Botany*, 2009.
- [14] T. Gibson and D. Waller, "Evolving Darwin's 'most wonderful' plant: ecological steps to a snaptrap", *New Phytologist*, 2009.
- [15] J. Greilhuber, T. Borsch, and K. Müller, "Smallest angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size", *Plant Biology*, 2006.
- [16] Y. Hoshi and K. Kondo, "A chromosome phylogeny of the *Droseraceae* by using CMA-DAPI fluorescent banding", *Cytologia*, 1998.
- [17] E. Ibarra-Laclette and V. Albert, "Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome", *BMC plant Biology*, 2011.
- [18] R. Jobson and R. Nielsen, "Adaptive evolution of cytochrome c oxidase: Infrastructure for a carnivorous plant radiation", *Proceedings of the National Academy of Sciences*, 2004.
- [19] H. Meimberg, P. Dittrich, G. Bringmann, J. Schlauer, and G. Heubl, "Molecular Phylogeny of *Caryophyllidae* s.l. Based on MatK Sequences with Special Emphasis on Carnivorous Taxa", *Plant Biology*, Vol. 2, No. 2, pp. 218–228, Mar. 2000.
- [20] X. Min, G. Butler, R. Storms, and A. Tsang, "TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences", *Nucleic acids research*, 2005.
- [21] J. Muller, "Fossil pollen records of extant angiosperms", *The Botanical Review*, 1981.
- [22] J. Pellicer, M. Fay, and I. Leitch, "The largest eukaryotic genome of them all?", *Botanical Journal of the Linnean Society*, 2010.
- [23] S. Poppinga, "Trap diversity and evolution in the family *Droseraceae*", *Plant signaling & behavior*, 2013.
- [24] T. Renner and C. Specht, "Molecular and functional evolution of class I chitinases for plant carnivory in the *Caryophyllales*", *Molecular biology and evolution*, 2012.
- [25] F. Rivadavia and K. Kondo, "Phylogeny of the sundews, *Drosera* (*Droseraceae*), based on chloroplast *rbcl* and nuclear 18S ribosomal DNA sequences", *American Journal of Botany*, 2003.
- [26] W. Rogers and J. Cruse-Sanders, "Development and characterization of microsatellite markers in *Sarracenia* L. (pitcher plant) species", *Conservation genetics resources*, 2010.
- [27] S. Rozen and H. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers.", *Methods in molecular biology*, Vol. 132, pp. 365–86, Jan. 2000.
- [28] M. Schulz, D. Zerbino, M. Vingron, and E. Birney, "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels", *Bioinformatics*, 2012.
- [29] W. Schulze and K. Sanggaard, "The protein composition of the digestive fluid from the Venus flytrap sheds light on prey digestion mechanisms", *Molecular & Cellular Proteomics*, 2012.
- [30] D. Sirová and J. Borovec, "Utricularia carnivory revisited: plants supply photosynthetic carbon to traps", *Journal of Experimental Botany*, 2010.
- [31] A. Srivastava, W. Rogers, and C. Breton, "Transcriptome analysis of *Sarracenia*, an insectivorous plant", *DNA research*, 2011.
- [32] D. Swarbreck and C. Wilks, "The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation", *Nucleic acids research*, 2008.
- [33] S. Wicke, G. Schneeweiss, K. Müller, and D. Quandt, "The evolution of the plastid chromosome in land plants: gene content, gene order, gene function", *Plant molecular biology*, 2011.
- [34] J. Wilhelm, A. Pingoud, and M. Hahn, "Realtime PCR-based method for the estimation of genome sizes", *Nucleic acids research*, 2003.
- [35] P. Zhulidov and E. Bogdanova, "Simple cDNA normalization using kamchatka crab duplex-specific nuclease", *Nucleic acids research*, 2004.
- [36] P. Zhulidov and E. Bogdanova, "A method for the preparation of normalized cDNA libraries enriched with full-length sequences", *Russian Journal of Bioorganic Chemistry*, 2005.
- [37] D. Zwickl and D. Hillis, "Increased taxon sampling greatly reduces phylogenetic error", *Systematic Biology*, 2002.

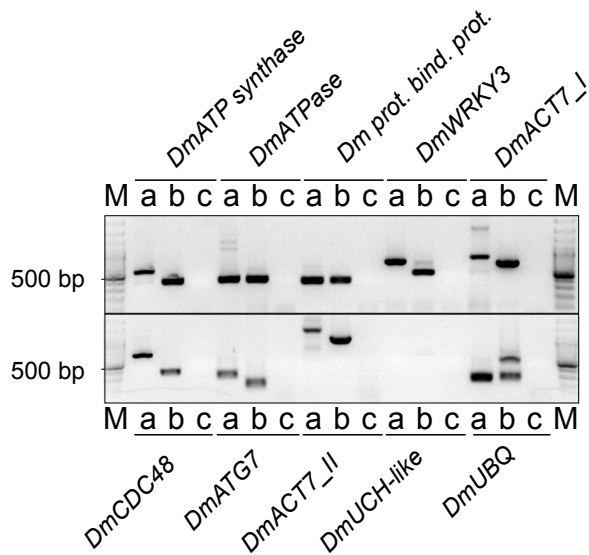


Figure 1: PCR assembly validation. Contigs assembled from 93 bp single-end reads were validated using standard PCR. A: genomic DNA, B: First-strand cDNA synthesis with reverse transcriptase, C: First-strand cDNA synthesis without reverse transcriptase. M: 100 bp O'GeneRuler. For primer and contig sequences see Supplementary file S1.

Table 1: Summary of BLASTx search results of *D. muscipula* transcriptome. From a total of 80,816 contigs, 42,656 have a RefSeq hit, corresponding to 17,047 unique protein entries. Total number and unique hits from a BLASTx against RefSeq entries for 8 other plant species is also presented. The percent of total unique proteins is based on the current number of RefSeq entries for the individual species.

Database	<i>D. muscipula</i> hits	Unique protein hits	% of total unique proteins
nr	42,656	17,047	–
Refseq/Ensembl			
<i>Arabidopsis thaliana</i>	41,422 (51.3%)	13,469	38.1% (13,469/35,378)
<i>Brachypodium distachyon</i>	39,962 (49.4%)	11,795	48.8% (11,795/24,689)
<i>Oryza sativa</i>	39,353 (48.7%)	11,506	40.1% (11,506/28,705)
<i>Physcomitrella patens</i>	34,084 (42.2%)	9,390	26.1% (9,390/35,936)
<i>Ricinus communis</i>	41,839 (51.7%)	12,279	39.1% (12,279/31,344)
<i>Vitis vinifera</i>	43,634 (53.9%)	12,837	53.8% (12,837/23,877)
<i>Zea mays</i>	35,229 (43.6%)	10,194	45.1% (10,194/22,588)
<i>Solanum lycopersicum</i>	42,489 (52.6%)	13,152	59.8% (13,152/26,408)

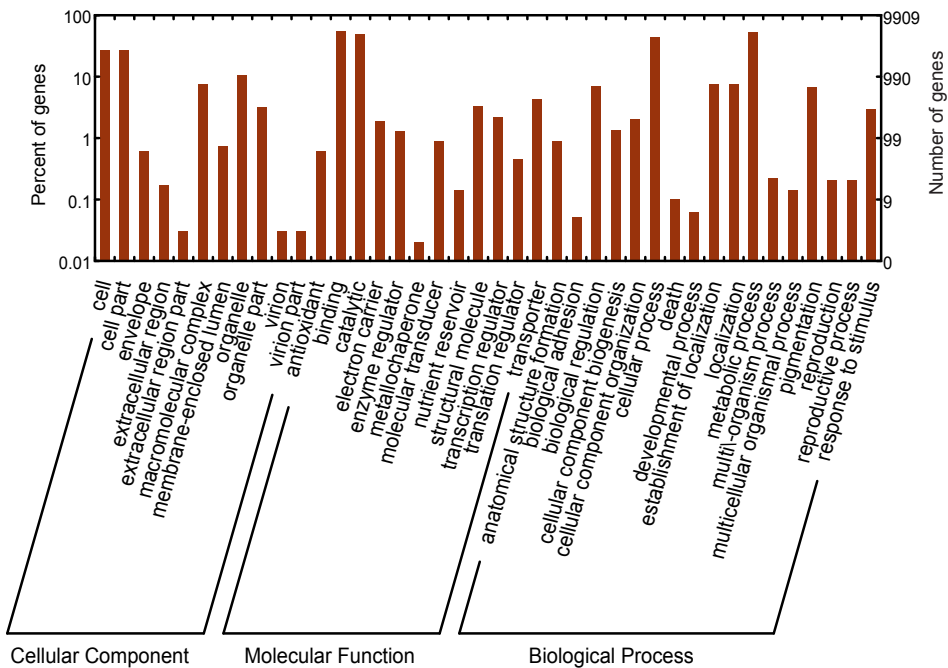


Figure 2: Gene Ontology (GO) categories of the unigenes. Distribution of the GO categories assigned to the *D. muscipula* transcriptome. Unique transcripts (unigenes) were annotated in three categories: cellular components, molecular functions, and biological processes.

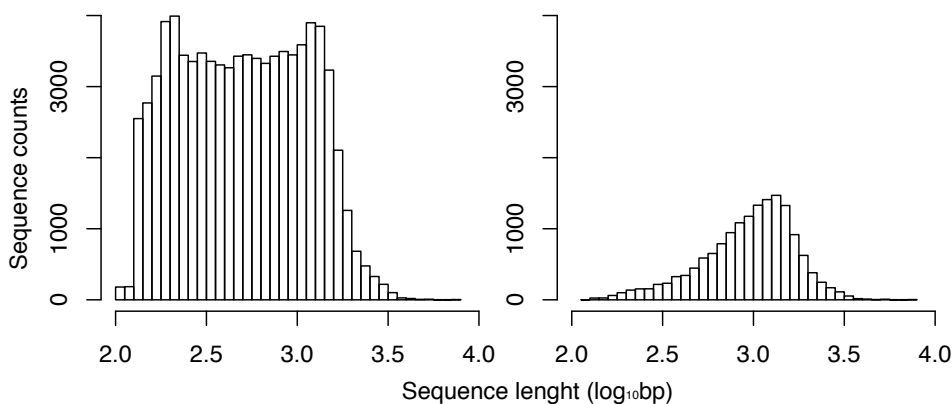


Figure 3: Contig size distribution of all contigs (left) and predicted full-length contigs (right).

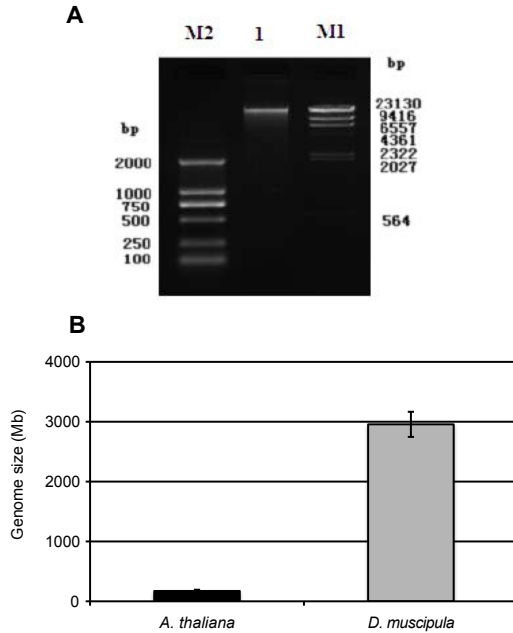


Figure 4: Genomic DNA purification and genome size estimate of *D. muscipula*. (A) Agarose gel showing a purified fraction of *D. muscipula* genomic DNA (1) using a modified CTAB procedure. M1: DNA ladder D2000 (Tiangen), M2: DNA ladder λ -Hind3 digest (Takara). (B) Genome size estimate of *D. muscipula* using a single-copy qPCR method with DmACT7 as amplicon. *A. thaliana* serves as a control, using ACTIN1 as amplicon.

Table 3: Summary of qPCR-based estimates of haploid genome sizes

Target	Product length (bp)	Calibration curve $y = mx + b$ (R^2)	Genome size estimate +/- SEM (Mbp)	n	1C +/- SEM (pg)
ACT1 (At2g37620) (<i>A. thaliana</i>)	116	-3.263X + 45.613 (0.995)	173 +/- 21	7	0.17 +/- 0.02
ACT7 (<i>D. muscipula</i>)	185	-3.323X + 36.6 (0.994)	2956 +/- 210	11	3.02 +/- 0.21

Part IV

Epilogue

Chapter 12

Summary and Perspectives

12.1 Summary

Overall, the work presented in this thesis provides an extensive collection of methods and tools for both analyzing transcriptomic and metagenomic sequencing data for comparative functional analysis and a concept of how to tap into biological resources for potential commercial use. The polar marine environment is a unique ground for identifying novel proteases. Here we identify numerous candidate sequences for further screening analysis which will be achieved with industrial partners. Furthermore, unusual organisms such as the Venus flytrap might prove to reveal novel features in proteolytic enzymes for industrial applications such as the food industry.

This thesis elucidates the use of Next Generation sequencing (NGS) analysis in uncovering functional descriptions of the Venus flytrap *Dionaea muscipula* transcriptome and environmental samples of the polar seas. The research field of metagenomics and RNA-seq analysis are introduced as concepts of how to access genomic information and identify transcripts. Furthermore, it is described how tapping into biological resources can be performed to identify novel proteases for commercial use. The methods section provides an overview of how to process and analyze transcriptome and metagenome sequencing data, the latter was illustrated with aid of the taxonomic annotation of a middle Pleistocene horse metagenome and the donkey genome annotation (Manuscript I).

The polar marine samples were collected during the Galathea III and LOMROG II polar expeditions and the DNA was extracted and sequenced. To my knowledge, no other dataset of the polar regions including multiple locations from the Arctic and Southern oceans at depth up to 4,300 m is

publicly available, making this dataset a unique resource for identifying the functional composition of the polar seas. The comparative functional analysis revealed a functional stratification along the water column but a lesser stratification between the arctic and southern oceans (Manuscript II). Furthermore, strategies for deep sea persistence were revealed. This unique dataset is a perfect subject for bioprospecting. 2,707 novel proteolytic enzyme sequences were identified by bioinformatic analysis (Manuscript III). A patent application was filed at DTU and we are currently negotiating the use of the study results and dataset with potential business partners. We anticipate that these sequences exhibit novel properties which can be used for commercial use. The collaboration of research institutions and companies is common. Despite the high interest in natural products, many large companies were not expanding their in-house natural products programs, but they were licensing in, or forming partnerships with small companies and universities that aid in discovery research [90]. However, the advances of metagenomics make collection of interesting dataset more cost-effective. Companies such as Verenium are actively collecting datasets keeping large in-house databases for screening tests¹. While collecting the samples the CBD has become the rule to follow. Although few companies were reported to by-pass these standards², companies do not consider genetic resources freely available. Companies regard benefit-sharing and compliance to the CBD as a necessary business practice. On the other hand, if national authorities set their demands of benefit-sharing too high, research opportunities can be lost. This becomes apparent with the attempt to research the Ikka column samples for bioprospecting. The samples could not be utilized due to limitations set by the Greenlandic government.

The transcriptome of the Venus flytrap was sequenced with a cost-effective approach using a normalized library of mixed tissues from trap and flowers (Manuscript IV). To expand the list of genome size estimates of members of the carnivorous orders, we present the first genome size estimate of a member of the sundew family in the order Caryophyllales. The genome was estimated to be 2,956 Mbp and exceeds the genome size of *Arabidopsis thaliana* with a genome size of 157 Mbp by a factor of 20. The genome size is even higher when compared to *Genlisea margaretae*, which exceeds the size 46-fold. In connection with this project we also attempted to sequence the genome of the Venus flytrap, which could not be assembled to contigs with reasonable length for further analysis. The assembly was affected by the high content of repetitive regions. The transcriptome provides insight into the molecular processes occurring in a Gb-sized carnivorous plant genome. The transcriptome data and the genome size estimate expands the knowledge of this very unique plant species. Future studies may address the identification of tissue

¹<http://www.verenium.com/ourworld.html>, accessed 13. November 2013.

²adopted from the United Nations University's report: Biological prospecting in Antarctica, http://www.ias.unu.edu/sub_page.aspx?catID=35&ddlID=20, accessed 20. November 2013.

and developmentally regulated genes by temporal and spatial sampling of tissues under different conditions, such as comparison of expressed genes under starvation and digestion. Furthermore, the data can be used as an annotative tool for the future *de novo* assembly of the Venus flytrap genome. A completed genome will facilitate the identification of novel proteases which are secreted during the digestion of its prey. This example illustrates the limitations of Next Generation sequencing data of larger genomes with high ploidity and repetitive sequences. Longer sequencing reads would provide extra information for scaffolding to the *de novo* assembly. Newer technologies such as Single-molecule real-time sequencing (Pacific Bio¹) are promising to aid in closing that gap by providing longer sequencing reads.

Lastly, I would like to give perspectives on some parts of the presented research.

2,707 novel sequences were identified from the marine gene catalogue. Currently, negotiations with industrial partners are ongoing. We are working on settling an agreement, after which some of the sequences will be enzymatically screened. Furthermore, the gene catalogue provides an abundant resource also for mining of other classes of enzymes such as hydrolases. In recent years, research efforts has been subjected to identify glycosyl hydrolases [3, 9, 28, 56, 68, 99, 169] due to the demand for finding new enzymes which can be used in cost-effective processes to break down biomass as a source of renewable energy. Therefore, the gene catalogue of the polar marine environment possesses value for other classes of enzymes besides proteases.

Metagenomics is a relative new research field and the advances in methods facilitating the analysis of large datasets is progressing rapidly. Improved methods for binning of contigs will help in downstream analysis. The principle of metagenomic species (MGS) for example has been shown to aid in single genome assembly of bacteria in metagenomes with the aid of canopy clustering methods (Nielsen *et al.*, unpublished). This is intriguing as it potentially leads to identification of novel organisms of the polar sea. The taxonomic annotation of the metagenomic samples revealed a high fraction of unassignable sequencing reads. Thus, extended databases with sequence information of hitherto unknown organisms can improve taxonomic annotation studies. Moreover, the MGS approach can be used to identify species that co-occur in a community or exclude each other. The same can be done for phages, plasmids and other genetic elements.

Moreover, we are planning to collaborate with researchers from University of Copenhagen to identify environmental DNA (eDNA) from the polar water samples. eDNA is free DNA in solution, in contrary to genomic DNA in cells. It has been shown that traces of fish DNA can be identified in fresh and sea water samples by sequencing eDNA [165]. After screening for fish sequences in the metagenomes, traces of fish DNA have already been found in the raw

¹Pacific Biosciences Introduces New Chemistry With Longer Read Lengths to Detect Novel Features in DNA Sequence and Advance Genome Studies of Large Organisms, available at <http://www.globenewswire.com/>, accessed 14. November 2013

sequencing data. This, however, has to be confirmed with fish specific PCR primers in the extracted DNA samples and in the filter remains. Such approaches might lead to mapping of not just bacterial communities but also multicellular organisms.

Bibliography

- [1] Adamec L. (1995). Ecological requirements and recent European distribution of the aquatic carnivorous plant *Aldrovanda vesiculosa* L.—A review. *Folia Geobotanica*. 99
- [2] Albert V.A., Williams S.E., and Chase M.W. (1992). Carnivorous plants: phylogeny and structural evolution. *Science*, 257(5076):1491–5. 99
- [3] Allgaier M., Reddy A., Park J.I., Ivanova N., D'haeseleer P., et al. (2010). Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. *PLoS one*, 5(1):e8812. 117
- [4] Altschul S. and Madden T. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 32, 36, 52
- [5] Amann R.L., Ludwig W., and Schleifer K.H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–69. 10
- [6] Arumugam M., Raes J., Pelletier E., Le Paslier D., Yamada T., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80. 51
- [7] Athauda S.B.P., Matsumoto K., Rajapakshe S., Kuribayashi M., Kojima M., et al. (2004). Enzymic and structural characterization of nepenthesin, a unique member of a novel subfamily of aspartic proteinases. *The Biochemical journal*, 381(Pt 1):295–306. 101
- [8] Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., et al. (2005). The Universal Protein Resource (UniProt). *Nucleic acids research*, 33(Database issue):D154–9. 36, 59
- [9] Banerjee G., Scott-Craig J.S., and Walton J.D. (2010). Improving Enzymes for Biomass Conversion: A Basic Research Perspective. *BioEnergy Research*, 3(1):82–92. 117
- [10] Barrett A., Woessner J., and Rawlings N. (2004). *Handbook of proteolytic enzymes*. Elsevier. 19
- [11] Barthlott W., Porembski S., Seine R., and Theisen I. (2007). *The Curious World of Carnivorous Plants: A Comprehensive Guide to Their Biology and Cultivation*. Timber Press. 99

- [12] Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., et al. (2004). The Pfam protein families database. *Nucleic acids research*, 32(Database issue):D138–41. 36, 38
- [13] Baum L.E. and Petrie T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563. 31
- [14] Bernal A. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic acids research*, 29(1):126–127. 11
- [15] Bessis A.S., Rondard P., Gaven F., Brabet I., Triballeau N., et al. (2002). Closure of the Venus flytrap module of mGlu8 receptor and the activation process: Insights from mutations converting antagonists into agonists. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17):11097–102. 99
- [16] Blencowe B.J., Ahmad S., and Lee L.J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & development*, 23(12):1379–86. 29
- [17] Bloom S. (1981). Similarity indices in community studies: potential pitfalls. *Marine Ecology Progress Series*, 5:125–128. 51
- [18] Boeckmann B. and Bairoch A. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*. 32
- [19] Bray J.R. and Curtis J.T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325. 51
- [20] Buchardt B., Seaman P., Stockmann G., Vous M., Wilken U., et al. (1997). Submarine columns of ikaite tufa. *Nature*, 390(6656):129–130. 5
- [21] Cameron K.M., Wurdack K.J., and Jobson R.W. (2002). Molecular evidence for the common origin of snap-traps among carnivorous plants. *American journal of botany*, 89(9):1503–9. 99, 100
- [22] Camon E., Magrane M., Barrell D., Lee V., Dimmer E., et al. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research*, 32(Database issue):D262–6. 35
- [23] Campbell M., Haas B., and Hamilton J. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, 7(327). 29
- [24] Carmack E.C., Aagaard K., Swift J.H., MacDonald R.W., McLaughlin F.A., et al. (1997). Changes in temperature and tracer distributions within the Arctic Ocean: results from the 1994 Arctic Ocean section. *Deep Sea Research Part II: Topical Studies in Oceanography*, 44(8):1487–1502. 56
- [25] Chain P.S.G., Grafham D.V., Fulton R.S., Fitzgerald M.G., Hostetler J., et al. (2009). Genomics. Genome project standards in a new era of sequencing. *Science (New York, N.Y.)*, 326(5950):236–7. 30
- [26] Chaisson M.J., Brinza D., and Pevzner P.A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research*, 19(2):336–46. 27, 28
- [27] Charuvaka A. and Rangwala H. (2011). Evaluation of short read metagenomic assembly. *BMC genomics*, 12 Suppl 2(Suppl 2):S8. 29

- [28] Chistoserdova L. (2010). Recent progress and new challenges in metagenomics for biotechnology. *Biotechnology letters*, 32(10):1351–9. 117
- [29] Church M.J., Ducklow H.W., and Karl D.M. (2004). Light dependence of [3H]leucine incorporation in the oligotrophic North Pacific ocean. *Applied and environmental microbiology*, 70(7):4079–87. 56
- [30] Cock P.J.A., Fields C.J., Goto N., Heuer M.L., and Rice P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71. 26
- [31] Compeau P.E.C., Pevzner P.A., and Tesler G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–91. 27
- [32] Coughlin Jr M.D. (1993). Using the Merck-INBio agreement to clarify the Convention on Biological Diversity. *Columbia Journal of Transnational Law*, 31:337. 17
- [33] Cropper A. (2009). Convention on Biological Diversity. *Environmental Conservation*, 20(04):364. 3, 17
- [34] Darwin C. (1875). *Insectivorous Plants*. John Murray, London. 99
- [35] Delcher A.L., Harmon D., Kasif S., White O., and Salzberg S.L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic acids research*, 27(23):4636–41. 32
- [36] Delille D. and Perret E. (1989). Influence of temperature on the growth potential of Southern polar marine bacteria. *Microbial ecology*, 18(2):117–23. 79
- [37] DeLong E.F., Preston C.M., Mincer T., Rich V., Hallam S.J., et al. (2006). Community genomics among stratified microbial assemblages in the ocean’s interior. *Science (New York, N.Y.)*, 311(5760):496–503. 56
- [38] Denoeud F., Aury J.M., Da Silva C., Noel B., Rogier O., et al. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome biology*, 9(12):R175. 29
- [39] Eddy S. (2004). What is dynamic programming? *Nature biotechnology*, 22(7):909–910. 32
- [40] Eddy S.R. (2004). What is a hidden Markov model? *Nature biotechnology*, 22(10):1315–6. 31, 32
- [41] Eddy S.R. (2011). Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195. 52
- [42] Edgar R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 32, 36
- [43] Eisner T. (1991). Chemical prospecting: a proposal for action. *Ecology, economics, and ethics: The broken circle*, pages 196–202. 16
- [44] Ellison A. and Gotelli N. (2009). Energetics and the evolution of carnivorous plants—Darwin’s ‘most wonderful plants in the world’. *Journal of Experimental Botany*, 60(1):19–42. 99
- [45] Ewing B. and Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3):186–94. 26
- [46] Ewing B., Hillier L., Wendl M.C., and Green P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8(3):175–185. 26

- [47] Falkowski P.G., Fenchel T., and Delong E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science*, 320(5879):1034–9. 10
- [48] Finn R.D., Mistry J., Schuster-Böckler B., Griffiths-Jones S., Hollich V., et al. (2006). Pfam: clans, web tools and services. *Nucleic acids research*, 34(Database issue):D247–51. 37
- [49] Finn R.D., Mistry J., Tate J., Coghill P., Heger A., et al. (2010). The Pfam protein families database. *Nucleic acids research*, 38(Database issue):D211–22. 37
- [50] Flicek P., Amode M., and Barrell D. (2012). Ensembl 2012. *Nucleic acids research*, 40(Database issue):84–90. 30, 32
- [51] Flusberg B.A., Webster D.R., Lee J.H., Travers K.J., Olivares E.C., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461–5. 9
- [52] Forterre Y., Skotheim J.M., Dumais J., and Mahadevan L. (2005). How the Venus flytrap snaps. *Nature*, 433(7024):421–5. 99
- [53] Friedman W. (2009). The meaning of Darwin's "abominable mystery". *American Journal of Botany*, 96(1):5–21. 99
- [54] Geikie J. (1895). The "challenger" expedition. *The Scottish Geographical Magazine*. 55
- [55] Gibson T. and Waller D. (2009). Evolving Darwin's 'most wonderful' plant: ecological steps to a snap-trap. *New Phytologist*, 183(3):575–587. 99
- [56] Gilbert J., Li L.L., Taghavi S., McCorkle S.M., Tringe S., et al. (2012). Bio-prospecting metagenomics for new glycoside hydrolases. *Methods in molecular biology (Clifton, N.J.)*, 908:141–51. 117
- [57] Giovannoni S.J., Rappe M.S., Vergin K.L., and Adair N.L. (1996). 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *Proceedings of the National Academy of Sciences*, 93(15):7979–7984. 56
- [58] Gnerre S., Lander E.S., Lindblad-Toh K., and Jaffe D.B. (2009). Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome biology*, 10(8):R88. 27
- [59] Gnerre S., Maccallum I., Przybylski D., Ribeiro F.J., Burton J.N., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–8. 27, 28
- [60] Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52. 28, 30
- [61] Haas B.J. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19):5654–5666. 29
- [62] Handelsman J., Tiedje J., Alvarez-Cohen L., Ashburner M., Cann I.K.O., et al. (2007). *The new science of metagenomics: revealing the secrets of our microbial planet*. The National Academies Press, Washington, DC. 10, 17, 18
- [63] Harris M., Clark J., and Ireland A. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32((Database issue)):D258–61. 35, 36

- [64] Hartmann T. (2008). The lost origin of chemical ecology in the late 19th century. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4541–6. 16
- [65] Heber S., Alekseyev M., Sze S.H., Tang H., and Pevzner P.A. (2002). Splicing graphs and EST assembly problem. *Bioinformatics*, 18(Suppl 1):S181–S188. 29
- [66] Hedstrom L. (2002). Serine protease mechanism and specificity. *Chemical reviews*, 102(12):4501–24. 19
- [67] Hemmings A. (2010). Does bioprospecting risk moral hazard for science in the Antarctic Treaty System? *Ethics in Science and Environmental Politics*, 10:5–12. 18
- [68] Hess M., Sczyrba A., Egan R., Kim T.W., Chokhawala H., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N. Y.)*, 331(6016):463–7. 117
- [69] Hewson I. and Steele J. (2006). Remarkable heterogeneity in meso- and bathypelagic bacterioplankton assemblage composition. *Limnology and Oceanography*, 51(3):1274–1283. 56
- [70] Höft R. (2001). *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK. 17
- [71] Huang W.E., Zhou J., Desai N., Antonopoulos D., Gilbert J.A., et al. (2012). From genomics to metagenomics. *Current Opinion in Biotechnology*, 23(1):72–76. 29
- [72] Hubbard T., Barker D., and Birney E. (2002). The Ensembl genome database project. *Nucleic acids research*. 30, 32
- [73] Hyatt D., Chen G.L., Locascio P.F., Land M.L., Larimer F.W., et al. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):119. 31, 32, 33
- [74] Jacobsen A., Hendriksen R.S., Aaresturp F.M., Ussery D.W., and Friis C. (2011). The *Salmonella enterica* pan-genome. *Microbial ecology*, 62(3):487–504. 36
- [75] Jannasch H. and Taylor C. (1984). Deep-sea microbiology. *Annual Reviews in Microbiology*. 55
- [76] Jaouadi B., Abdelmalek B., Jaouadi N., and Bejar S. (2011). *The bioengineering and industrial applications of bacterial alkaline proteases: the case of SAPB and KERAB*. InTech. 21, 79
- [77] Jiang H. and Wong W.H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics (Oxford, England)*, 25(8):1026–32. 29
- [78] Johannessen O.M., Muench R.D., and Overland J.E., editors (1994). *The Polar Oceans and Their Role in Shaping the Global Environment*, volume 85 of *Geophysical Monograph Series*. American Geophysical Union, Washington, D. C. 56
- [79] Jollivet D. (1996). Specific and genetic diversity at deep-sea hydrothermal vents: an overview. *Biodiversity & Conservation*. 55
- [80] Kageyama T. (2002). Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. *Cellular and molecular life sciences : CMLS*, 59(2):288–306. 101
- [81] Kahvejian A., Quackenbush J., and Thompson J.F. (2008). What would you do if you could sequence everything? *Nature biotechnology*, 26(10):1125–33. 7, 13

- [82] Kanehisa M. and Goto S. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34. 36
- [83] Katoh K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066. 52
- [84] Kelley D.R., Schatz M.C., and Salzberg S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116. 26
- [85] Kerkhof L.J. and Goodman R.M. (2009). Ocean microbial metagenomics. *Deep Sea Research Part II: Topical Studies in Oceanography*, 56(19-20):1824–1829. 11
- [86] Kirchman D.L., Morán X.A.G., and Ducklow H. (2009). Microbial growth in the polar oceans - role of temperature and potential impact of climate change. *Nature reviews. Microbiology*, 7(6):451–9. 56
- [87] Kirk O., Borchert T., and Fuglsang C. (2002). Industrial enzyme applications. *Current opinion in biotechnology*, 13(4):345–51. 21
- [88] Knight R., Jansson J., Field D., Fierer N., Desai N., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature biotechnology*, 30(6):513–20. 10
- [89] Kruskal W.H. and Wallis W.A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621. 52
- [90] Laird S., Wynberg R., and Johnston S. (2006). Recent Trends in the Biological Prospecting. *Antarctic Treaty Consultative Meeting 2006*, 17(2):61–70. 116
- [91] Lamed Y. (1996). A Venus Flytrap Mechanism for Activation and Desensitization of alpha -Amino-3-hydroxy-5-methyl-4-isoxazole Propionic Acid Receptors. *Journal of Biological Chemistry*, 271(26):15299–15302. 99
- [92] Langmead B. and Salzberg S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9. 38
- [93] Law B.A. (2009). Enzymes of psychrotrophic bacteria and their effects on milk and milk products. *Journal of Dairy Research*, 46(03):573. 79
- [94] Le H.S., Schulz M.H., McCauley B.M., Hinman V.F., and Bar-Joseph Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic acids research*, 41(10):e109. 26
- [95] Le Chatelier E., Nielsen T., Qin J., Prifti E., Hildebrand F., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–6. 35
- [96] Lewin A., Wentzel A., and Valla S. (2012). Metagenomics of microbial life in extreme temperature environments. *Current opinion in biotechnology*, 24(3):516–25. 12
- [97] Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60. 27, 38, 50, 51
- [98] Li H., Handsaker B., Wysoker A., and Fennell T. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9. 50
- [99] Li L., McCorkle S., and Monchy S. (2009). Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnology for Biofuels*, 2(10). 117
- [100] Li Q., Yi L., Marek P., and Iverson B.L. (2013). Commercial proteases: present and future. *FEBS letters*, 587(8):1155–63. 19

- [101] Li W. and Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–9. 49
- [102] Liang C., Mao L., Ware D., and Stein L. (2009). Evidence-based gene predictions in plant genomes. *Genome research*, 19(10):1912–23. 33
- [103] Liu L., Li Y., Li S., Hu N., He Y., et al. (2012). Comparison of next-generation sequencing systems. *BioMed Research*. 9
- [104] Loman N.J., Misra R.V., Dallman T.J., Constantinidou C., Gharbia S.E., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5):434–9. 9
- [105] Luo R., Liu B., Xie Y., Li Z., Huang W., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18. 28
- [106] MacLean D., Jones J.D.G., and Studholme D.J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(2):96–97. 27
- [107] Martin J.A. and Wang Z. (2011). Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82. 14
- [108] Marx J.C., Collins T., D'Amico S., Feller G., and Gerday C. (2007). Cold-adapted enzymes from marine Antarctic microorganisms. *Marine biotechnology (New York, N.Y.)*, 9(3):293–304. 79
- [109] Mende D.R., Waller A.S., Sunagawa S., Järvelin A.I., Chan M.M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one*, 7(2):e31386. 29
- [110] Metzker M. (2009). Sequencing technologies—the next generation. *Nature reviews. Genetics*. 7, 9
- [111] Mgbeoji I. (2005). *Global biopiracy: patents, plants, and indigenous knowledge*. UBC Press. 16
- [112] Min X., Butler G., Storms R., and Tsang A. (2005). TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences. *Nucleic acids research*, 33(Web Server issue):W669–72. 30
- [113] Moran J., Clarke C., and Hawkins B. (2003). From carnivore to detritivore? Isotopic evidence for leaf litter utilization by the tropical pitcher plant *Nepenthes ampullaria*. *International Journal of Plant Sciences*, 164(4):635–639. 101
- [114] Mortazavi A., Williams B.A., McCue K., Schaeffer L., and Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8. 29
- [115] Muller J., Szklarczyk D., Julien P., Letunic I., Roth A., et al. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic acids research*, 38(Database issue):D190–5. 36, 37
- [116] Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9. 13
- [117] Namiki T., Hachiya T., Tanaka H., and Sakakibara Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155. 28, 29

- [118] Nelson K.E., editor (2011). *Metagenomics of the Human Body*. Springer New York, New York, NY. 10
- [119] Ogawa H. and Tanoue E. (2003). Dissolved Organic Matter in Oceanic Waters. *Journal of Oceanography*, 59(2):129–147. 56
- [120] Orlando L., Ginolhac A., Zhang G., Froese D., Albrechtsen A., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74–8. 39
- [121] Ozsolak F. and Milos P.M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, 12(2):87–98. 13
- [122] Parthasarathy H., Hill E., and MacCallum C. (2007). Global ocean sampling collection. *PLoS biology*, 5(3):e83. 55
- [123] Pati A., Ivanova N.N., Mikhailova N., Ovchinnikova G., Hooper S.D., et al. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Methods*, 7(6):455–457. 33
- [124] Peng Y., Leung H.C.M., Yiu S.M., and Chin F.Y.L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics (Oxford, England)*, 27(13):i94–101. 27, 28, 29
- [125] Petersen T.N., Brunak S.r., von Heijne G., and Nielsen H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–6. 52
- [126] Pevzner P.A., Tang H., and Waterman M.S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–53. 29
- [127] Pignatelli M. and Moya A. (2011). Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS one*, 6(5):e19984. 29
- [128] Polaina J. and MacCabe A. (2007). *Industrial enzymes: structure, function and applications*, volume 96. Springer. 19, 20
- [129] Porter K. and Feig Y. (1980). The use of DAPI for identification and enumeration of bacteria and blue-green algae. *Limnology and oceanography*, 25:943. 55
- [130] Pruitt K., Tatusova T., and Maglott D. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue):D61–5. 30, 32
- [131] Qin J., Li R., Raes J., and Arumugam M. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65. 51
- [132] Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13:341. 9
- [133] Ram R.J., VerBerkmoes N.C., Thelen M.P., Tyson G.W., Baker B.J., et al. (2005). Community Proteomics of a Natural Microbial Biofilm. *Science*, 308(5730):1915–1920. 10
- [134] Rapaport F., Khanin R., Liang Y., Pirun M., Krek A., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9):R95. 13

- [135] Rawlings N.D., Barrett A.J., and Bateman A. (2010). MEROPS: the peptidase database. *Nucleic acids research*, 38(Database issue):D227–33. 20, 21
- [136] Rawlings N.D., Barrett A.J., and Bateman A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 40(Database issue):D343–50. 20, 21, 52
- [137] Roberts A., Pimentel H., Trapnell C., and Pachter L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)*, 27(17):2325–9. 28
- [138] Robertson G., Schein J., Chiu R., Corbett R., Field M., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature methods*, 7(11):909–12. 28
- [139] Rounsley S., Marri P.R., Yu Y., He R., Sisneros N., et al. (2009). De Novo Next Generation Sequencing of Plant Genomes. *Rice*, 2(1):35–43. 30
- [140] Rusch D., Halpern A., and Sutton G. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*. 55, 79
- [141] Rusch D.B., Martiny A.C., Dupont C.L., Halpern A.L., and Venter J.C. (2010). Characterization of Prochlorococcus clades from iron-depleted oceanic regions. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37):16184–9. 29
- [142] Sayers E.W., Barrett T., Benson D.A., Bolton E., Bryant S.H., et al. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 39(Database issue):D38–51. 32
- [143] Schmidt M., Priemé A., and Stougaard P. (2006). Bacterial diversity in permanently cold and alkaline ikaite columns from Greenland. *Extremophiles : life under extreme conditions*, 10(6):551–62. 5
- [144] Schmieder R. and Edwards R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6):863–4. 26
- [145] Schulz M.M.H., Zerbino D.R.D., Vingron M., and Birney E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8):1086–92. 28, 30
- [146] Schulze W. and Sanggaard K. (2012). The protein composition of the digestive fluid from the Venus flytrap sheds light on prey digestion mechanisms. *Molecular & Cellular Proteomics*, 11(11):1306–19. 101
- [147] Schwab D., Simmons E., and Scala J. (1969). Fine structure changes during function of the digestive gland of Venus’s-flytrap. *American Journal of Botany*, 56(1):88–100. 99
- [148] Schweikert G., Behr J., and Zien A. (2009). mGene. web: a web service for accurate computational gene finding. *Nucleic acids research*, 37(Web Server issue):W312–6. 31, 32
- [149] Schweikert G. and Institutes M. (2009). mGene: A novel discriminative gene finding system. *Genome Research*, 19(11):2133–2143. 31, 32
- [150] Seshadri R., Kravitz S.A., Smarr L., Gilna P., and Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS biology*, 5(3):e75. 11, 52
- [151] Shendure J. and Ji H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45. 7, 9

- [152] Siezen R.J. and Kleerebezem M. (2011). The human gut microbiome: are we our enterotypes? *Microbial biotechnology*, 4(5):550–3. 35
- [153] Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., et al. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23. 27, 28
- [154] Smith C. and Baco A. (2003). Ecology of whale falls at the deep-sea floor. *Oceanography and Marine Biology: an Annual Review*, 41:311–354. 55
- [155] Sonnhammer E., Eddy S., and Durbin R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–20. 36
- [156] Spurrier J.D. (2003). On the null distribution of the Kruskal–Wallis statistic. *Journal of Nonparametric Statistics*, 15(6):685–691. 52
- [157] Stanke M., Diekhans M., Baertsch R., and Haussler D. (2008). Using native and syn-tenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)*, 24(5):637–44. 32, 33
- [158] Stanke M., Steinkamp R., Waack S., and Morgenstern B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research*, 32(Web Server issue):W309–12. 32, 33
- [159] Sul W.J., Oliver T.A., Ducklow H.W., Amaral-Zettler L.A., and Sogin M.L. (2013). Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6):2342–7. 56
- [160] Sultan M., Schulz M.H., Richard H., Magen A., Klingenhoff A., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (New York, N.Y.)*, 321(5891):956–60. 29
- [161] Tatusov R.L. (1997). A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637. 36
- [162] Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):41. 36
- [163] Teeling H., Fuchs B.M., Becher D., Klockow C., Gardebrecht A., et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science (New York, N.Y.)*, 336(6081):608–11. 56
- [164] Temperton B. and Giovannoni S. (2012). Metagenomics: microbial diversity through a scratched lens. *Current Opinion in Microbiology*. 29
- [165] Thomsen P.F., Kielgast J., Iversen L.L.n., Møller P.R., Rasmussen M., et al. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PloS one*, 7(8):e41732. 117
- [166] Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5. 29
- [167] Tringe S.G., von Mering C., Kobayashi A., Salamov A.A., Chen K., et al. (2005). Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7. 10, 35
- [168] Ussery D.W., Wassenaar T.M., and Borini S. (2009). *Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists*. Springer. 10

- [169] van der Lelie D., Taghavi S., McCorkle S.M., Li L.L., Malfatti S.A., et al. (2012). The metagenome of an anaerobic microbial community decomposing poplar wood chips. *PLoS one*, 7(5):e36740. 117
- [170] Vazquez S., Rios Merino L., MacCormack W., and Fraile E. (1995). Protease-producing psychrotrophic bacteria isolated from Antarctica. *Polar Biology*, 15(2). 79
- [171] Velculescu V.E., Zhang L., Zhou W., Vogelstein J., Basrai M.A., et al. (1997). Characterization of the Yeast Transcriptome. *Cell*, 88(2):243–251. 13
- [172] Venter J., Remington K., and Heidelberg J. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(April):66–74. 10
- [173] von Bubnoff A. (2008). Next-generation sequencing: the race is on. *Cell*, 132(5):721–3. 7
- [174] Wang X., Wang H., Wang J., Sun R., Wu J., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature genetics*, 43(10):1035–9. 30
- [175] Wang Z., Gerstein M., and Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63. 29
- [176] Ward O.P. (2011). 3.49 - Proteases. In E.i.C.M. Moo-Young, editor, *Comprehensive Biotechnology (Second Edition)*, pages 571–582. Academic Press, Burlington, second edition. 21, 79
- [177] Webb E.C., Biology I.U.o.B., and Molecular (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Ed. 6. Academic Press Inc. 19
- [178] Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., et al. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 35(Database issue):D5–12. 32
- [179] Whitman W.B. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583. 10, 55, 56
- [180] Wilhelm L.J., Tripp H.J., Givan S.A., Smith D.P., and Giovannoni S.J. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biology direct*, 2:27. 29
- [181] Williams S.E. and Bennett A.B. (1982). Leaf closure in the venus flytrap: an Acid growth response. *Science (New York, N.Y.)*, 218(4577):1120–2. 99
- [182] Wooley J.C., Godzik A., and Friedberg I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667. 11, 27, 29
- [183] Worthington L. (1968). *Genesis and evolution of water masses*. Woods Hole Oceanographic Institution. 56
- [184] Xiong H., Song L., Xu Y., Tsoi M.Y., Dobretsov S., et al. (2007). Characterization of proteolytic bacteria from the Aleutian deep-sea and their proteases. *Journal of industrial microbiology & biotechnology*, 34(1):63–71. 5
- [185] Xu J., Ji P., Wang B., Zhao L., Wang J., et al. (2013). Transcriptome Sequencing and Analysis of Wild Amur Ide (*Leuciscus waleckii*) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress. *PLoS one*, 8(4):e59703. 30

- [186] Yakimov M.M., Cono V.L., Smedile F., DeLuca T.H., Juárez S., et al. (2011). Contribution of crenarchaeal autotrophic ammonia oxidizers to the dark primary production in Tyrrhenian deep waters (Central Mediterranean Sea). *The ISME journal*, 5(6):945–61. 56
- [187] Yassour M., Kaplan T., Fraser H.B., Levin J.Z., Pfiffner J., et al. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3264–9. 29
- [188] Zdobnov E.M. and Apweiler R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848. 38
- [189] Zeng R., Zhang R., Zhao J., and Lin N. (2003). Cold-active serine alkaline protease from the psychrophilic bacterium *Pseudomonas* strain DY-A: enzyme purification and characterization. *Extremophiles : life under extreme conditions*, 7(4):335–7. 79
- [190] Zerbino D.R. and Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9. 28
- [191] Zhu D., Wu Q., and Wang N. (2011). 3.02 - Industrial Enzymes. In E.i.C.M. Moo-Young, editor, *Comprehensive Biotechnology (Second Edition)*, pages 3–13. Academic Press, Burlington, second edi edition. 21
- [192] Zhu W., Lomsadze A., and Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132. 31, 32, 33
- [193] Zwirgmaier K., Jardillier L., Ostrowski M., Mazard S., Garczarek L., et al. (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environmental microbiology*, 10(1):147–61. 56

**Supplementary information:
Manuscript IV**

Supplementary file S1

>DmATP synthase Locus_396_Transcript_1/2_Confidence_1.000_Length_426
ATP synthase GI:109066521
atacgagatgtagaatcaagaggatcaacag**caggatagataccgagctc**agatatctgccgggacaac
acggttgtggcatccaagtgagcaaaggtggttgaggagcgggatccgtcaaatcatcagcaggcaca
taaatagcttgacggatgtgatggaacctttttcgtagttgtgatacgcctctgaaggcctccaagg
tcagtagcgggtaggctggtacccccacagcagaaggaatacgtccaagcaaaagcagacacctctgag
ttagcttgggtaaagcggaaaatattgtcaataaatagcagcacgtcttgtccctctgcatcacggaaa
tgttcagccacagtcgaagccagtaagac**caacacggggcagcagca**ccagggggctcattcatttgaccg
tagacaagagcg

F: caggatagataccgagctc
R: tgctcgtgcccggtgtt

Amplicon: 359 bp

>DmATPase - Locus_14083_Transcript_1/2_Confidence_1.000_Length_1605
ATPase gi|108708096
AAGCAGTGGTATCAACGCAGAGTACGGGGGCTTGGAAAGTGACGGACTTGAGTCATTGTGCATCAAGG
AGCAGACTGACAAATGAAAATGCTGAGAAGTTGTTGGATGGGCTCTAAGCTATCATTGATGCAGA
ACCAAAATGCGGAGTTAGAGGAGAAATAGTATTGTCTAGTGAGAGCCTCAGTATGGGATAGAAATCT
TACAGGCTATTCAAATGAGTCCAAAAGCTTAAAGAAGACACTGAAGGATGTTGTAACCGAAAATGAGT
TTGAGAAAAGGCTTCTGGCAGATGTTATTCACCCAGTGACATTGGGGTTACATTTGATGATATTGGTG
CCCTTGAGAATGTGAAGGATACATTGAAGGAGTTGGTGATGCTGCCATTGCAGAGGCCGGAACCTTTCT
GCAAGGGACAGTTAGCTAAGCCTTGCAAGGGCATACTTCTGTTTGGCCCTCCTGGCCTGGAAGACTA
TGCTTGCAAAAAGCTGTTGCAACAGAAGCCGGTGCGAACTTTATAAAATTTCCATGTCAAGCATCACAT
CTAAGTGGTTTGGTGAGGGTGAGAAATATGTGAAGGCTGTCTCTCTGGCAAGCAAGATTTCCCCTA
GTGTTGTGTTTGTGGATGAGGTTGATAGTATGCTCGGTGGAAGGGAACCCCTGGAGAGCATGAGGCCA
TGCGTAAAATGAAAATGAATTTATGGTGAATTTGGGATGGGCTGCGGACTAAAGATACAGAAAGAGTCC
TTGTAAGTGCAGCCACTAATAGGCCTTTGACCTTGATGAAGCGGTCATTAGAAGATTGCCCGGTAGGT
TAATGGTTAACTTGCCAGATGCTCCAAATAGAACTAAGATTCTTAAGGTGATATTGGCAAAAAGAAGAA
TATCTCATGATGTAGATTTAGATGCAGTTGGAAGCATGACCGAGGGATGTTCTGGGAGTGACCTCAAGA
ATCTTTGTGTTGCTGCTGCCACCGCCCTATCAGGGAGATTCTAGAAAAGGAAAAAAGGAGGCTGAAG
CTGCTGTGGCTGAAGGTAGACCCGCGCCACCTCTCAGTGGGAGTGCCGATATCCGGCCTCTAACATGG
ATGACTTCAAACACGCACATGAGCAGGTTTGTGCAAGTGTATCATCGGAGT**CCCATAACATGACCGAGC**
TTCAACAATGGAATGAACATATACGGCAGGGCGGCTCTAGGAAGAAGCGGCTCTAGTTACTTTATGT
AAATGCTTAATATTTCTAGTTATGTGAAAATGTATGTTCTCGGTGGTTCGTTTTGTTTTGGCCCGTATG
TGATGCCGACGGCTCTGCTTAGTGGAAATTTTTGTCATGTTTTTAGGAGGTTTGATACAGGTAAG
CGTGAGGCCCTTTTGTGTTAGGTTTGGCTCACCCCTGTATGAAGGTAAGTGTGTTGTTGCTGGGGGTATAA
GTTAGTTGAGAGGTATAAAGTTATAAAAATAGCGAGGACGGGA**AAGGAGGGCGTGTAGGAACT**GTCATAT
TGCCATTCTTTATTTTCAGATGTTAATAATATCTTCAGTTCGTCGTTTAAAAAAGTAC
TCTGCGTTGATACCACTG

F: CCCATAACATGACCGAGCTT
R: AGTTCCTACACCGCCTCTT

Amplicon 356 bp

>Dm prot. bind. prot.
Locus_17081_Transcript_1/2_Confidence_1.000_Length_2089 protein
binding protein, putative GI:25558542
ATCTAAGCAGTGGTATCAACGCAGAGTACGGGGAGCTGTGTGGACAAGGGTGATAGAAAGGTTGAGGGT
AACAGATTTACTGGGTCACATCATCGTTGGCACGATAATGAAGTAGGACCCTCAGAAGCAATTGATTTGT
ACGAAGCTCGCAGTGCAGTCACAAGAGTTTTTCTGCTAGAAATGGAGGAGGGCTCTTCATTTTCAGGTC

GGTAGTTGTGTTGCTACAGTTTCCGGAAC TCGTAAAAGGAAAAGCAGATGGGACCAGCCGTCGGATTCA
GAATTTCCCTTGCCAGGGACAAAAGGCGTTGCCAATTTTGTGCCAAAATTTTGATTCAGTCTACACCCCT
GAGACGGTCAAGGTTAGACTTCGTGCATATAAACAGAGCAAGGCATGTGGAGAAGAGTTCTACTGATTTGT
TCTGACGAGCAAAC TGAATATTGTACGACAGATAATGGAGTGATGGGCCCTCAGGGAGATGCTCCTCCT
GGATTTTCTTCCCATTTTCCCTGGATTCTCTCTCTCTCCTCCTCCTCCTCCTGGGTTTTCTTCCCT
CTTTGTGGTTCCCATCTTCAGTCATTTGTTGGCTCAACTGTGACTCACATCCCCAACAGAACAGAAAAG
CAGTTGCAATGTTGTAGCCCTTTCGGTGTCTCTTCAGGGCAGCTGCAACG**GAGGTTCAATTCTCGCTTG**
CCTGTAGCATATGGAATTCATTTTCAGCAATCCAGCAATTTGGGGCACCCCAACGGGGACACCTTGAT
GGTTGGGTTATTGCTCCAGGTATACCTTTTACCCTTTCCCCCTTGGCCACTTACCCCGTGGGTGC
AACAAAGGAGGTCAACAAGATTCTGGTGAATGACTGCAGAAAGGGTGAAGGGAAGTGGTAGACAAAATG
CCACAAGACCAC TGCACACATGTTTCTTATCGAGCAGATCAAGGCATGCCATGCGCCTCAGGTTGACT
TGTATAGATGTGGTTGCTGCCAG**TACATCCTTTCATCGGCCTCT**GCAGCAGGGAGGAGGCACCTCTTCC
AGCCTGGGAAGGAGTACTTTAGGCAGCAGAAGTGGAGGAACTCAAACAAAGGCCCGCTGGCTGAGA
AGGAACGACTGCGGATTCAAAGGAACTATCCCCAGAATGGGGTCTACGAGGGTGAATTTCCAGCGAG
CAGAATGTCAGTCTATGGAGTATATAAATTATCATGTAGACTGTGTTGGAATTTCTCGATAGCATCGA
ACCGTATCAAGAATACGACATTGATTTAACAATTTCTCCAGGGTTCAGAAAAAAGATTGAATTTACAT
TTAGATTATAGGAGAGTTGAGGATCTTCTGGATCTTGTACAGCCACCATAATTTCTTTTGTACATTGA
CGTTGGGGTACACCTTCATCTGAATTTCTGTTTTGTTGTAGTTTCGTTATACATTAGAGACATAATT
ATATGTGAATTTTAAAGAGAGATGAAGTTGACCAACCTTGTGGTGCCTCACCAACAAGAAAGGTGAGGT
TATATATTTGAACCTCTCACAGTAGCAAAACAATAAAATTTCCAAATTTCTATTTACCAGTAGTTTAT
TAAAACCAAAACCTTCTAGTAAGGAACATATGAAGTTTTGTCATGTAGAGCAGCACTCATCCGGTTGAGT
ACCGGCGACAAGAATCTGAAGCCAATTTATTGCCAGAAAAGCAGTCTTCTGCGGTTTATGATCTGGC
ATTTGCAGGATGTGAGAGAGCATGTTTCAAGTTTGGTTGCAGAACCATCAACATGTGGATCAGAAATTTCTA
TCAATGCTTAGCCTTTGAGGACATTTGTGTTAACTTGTTCACGATTTGTGACAGCAGGATTGAGCTC
GGTAAATTTGAGTACATTTGACTTTGAGCCAGGCCTTTTTAACTGATATGATTTATTGTTTACTGTTTT
TTTTGGCCATTTCAATCTAACATGACACCCCTTTTAGCCGTAATCCAGTTGTTTTTTCTGTGTAAGCTT
CCAAATTTCCACAACAAGTGAAGGATCCTTCTTGATAAGCTCACCATAAATCATCTCTGTTATAAAT
TGTCATGTTATCGGTGAT

F: GAGGTTCAATTCTCGCTTGC
R: GAGGCCGATGAAAGGATGTA

Amplicon : 338 bp

>DmWRKY3 Locus_5911_Transcript_3/3_Confidence_0.600_Length_1263 WRKY
DNA-binding protein 3 GI:15227612
TCCACGTCCTATGGTGCAGCCACCACCACAGTGTGAAATGGCTCAAATGGCCGCTCCTTCAAACCTTAGT
CCCTAAGGTTGGTGGAGAAGATCCTAAAACCTTCAGCAACTTCGGGTAATGCAGATAGACCCCTCCTACGA
TGGGTATAACTGGAGAAAATATGGTCAAAGCAGGTCAAAGGAAGCGAATACCCGAGAAGCTACTACAA
GTGACGCATCCAAACTGTCCAGTTAAAAGAAGGTAAAAGGTCGTTAGATGGGCAAAATAGCAGAAAT
TGCTTACAAGGGAGAACAACAACCATCCAAAGCCACAGCCCCCAAGCGCAGTTCTTCGGGAGTGAAGG
ACAAGGTTCAAGTGGCTGATGAGGTAGTCCAGGATCAAGATGGAAC TGCCTGACTGGCACTGGCACTGCCAC
TAATACCAAGTGGAAATATTGGTATTGTCAATGCAACTACTGAAGCTTTTGAAGGGCGATTAGAGAACCA
AAATGAAGTAGGATTGTCGACACAGTCAACTCATTCAAACAAGGCCGATTTTGTGCCTTTTGTATCTCT
TGCTGTAGCAATGGAGATGCTGATACTTGTGGTGAAGCACTGATTTTGAAGAAGGTAGCAGGGGATT
GGACGTCGATAATGATGAACCAAAAAGCAAGAAGAGGAGAAAAGATGGTCAAACAATGAAGCAGGACC
GAGCGGAGATGGTGTGCAAGTGAAGATCCTCCTCGTCATCTTCAAGTGC AAAGCACCACGGAACCTGA
GAGTTTAGGGGACGG**CTTTCGCTGGAGAAAATATGGCCAG**AAGGTCGTTAAAGGAAACCCGTATCCTAG
AAGTTACTACAGATGCACGAGCCTCAAATGCAACGTGCGAAAGCATGTAGAAAGAGCATCCGATGATCC
AAGATCATTCATCACCACGTACGAGGGGAAACACAACCACGAGATGCCATGAAAAGTACAAAATTCAGC
GGCGCCCTCCGAGCCAGATTCATCACAGCTCCTTTCTACAAGGACAAGAAGTGATTGACCTCACGGTG
ACTACTACATGTTTAAACCACTAAACAGTAAACACACCCCTGATGCCAGTCTTTTAAATCATTATATAGTT
TTTGTGGTGATATAGAAGCCTAATCAGCAAGTTCTCACTAGTCTGATTTATATATCATCACCACCTGAAA
TTCACTGTTTATCATCCTGTTTCG**CTACTAGAATCGTAAATTTGCTTAG**CTTATGCAAAAAAAAAAAAAA
GTACTCTGCGTTGATACCAC T

F: CTTTCGCTGGAGAAAATATGGCCAG
R: CTACTAGAATCGTAAATTTGCTT

Amplicon: 446 bp

>DmACT7_I 1394 bp

TGCTCATCTCACTCTGCAGGTATATAGAGAATGGCCGATGCTGAGGAGATTCAACCTCTTGTCTGTGAC
AATGGAACCTGGTATGGTGAAGGCTGGGTTTGTGGCGATGATGCTCCTAGGGCAGTGTTCAGTATAT
GTTGGCGTCCAGGCACACAGGTGTGATGGTTGGTATGGGACAGAAGGATGCTTATGTGGGTGATGAA
GCTCAATCTAAAAGAGGTATCCTTACCTTGAATACCCATTGAGCATGGCATTGTGAGCAACTGGGAT
GACATGGAGAAGATCTGGCATCACACTTCTACAACGAGCTCCGTGTTGCTCCTGAGGAGCATCCGGTG
CTTCTAACTGAGGCTCCTCTCAACCTAAGGCAACAGGGAAAAGATGACTCAAATCATGTTTGAGACA
TTCAATGTCCCTGCCATGTATGTTGCTATCCAGGCTGTTCTTTCTCTATGCCAGTGGTTCGTACAACG
GGTATCGTGTGGACTCTGGTGTGGTGTGAGTCACACTGTCCCATTTATGAAGGTTATGCACTTCCC
CATGCTATCCTTCGGCTGGACCTTGCTGGCCGCGACCTCACTGATTCTTATGAAGATTCTTACCAG
AGGGCTACATGTTACAACCA**CTGCTGAACGGGAAATTGTT**CGCGACATCAAGGAGAAGCTTGCATAT
GTAGCTCTGACTATGAGCAGGAGCTGGAACTGCCAAGAGCAGCAAGTTATTACCATAGGGGCTGAGA
GGTTCAGATGCCCTGAAGTCTCTTCCAGCCTCTTTGATTGGGATGGAAGCTGCTGGCATTATGAGA
CAACCTACAATTCATCATGAAGTGCACGTTGATATCAGGAAGGACTGTATGGTAACATCGTGCTTA
GTGGTGGTTCTACTATGTTCCCTGGCATTGCAGACAGGATGAGCAAGGAAATCACAGCACTTGCTCCAA
GCAGCATGAAGATCAAGGTGGTTGCTCCTCCAGAGAGGAAATACAGTGTCTGGATTGGAGGATCAATCC
TTGCATCTCTCAGCACCTTCCAACAGATGTGGATTCCAAGGGCAGTACGATGAGTCTGGTCCATCCA
TTGTCCACAGGAAATGCTTCTAAGCTCTACAGGATGCTTCGAGGGTGAGAGTCCAATATTTCTTTAGT
TGCTTGTGTGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCAAGTCA
TGGAAGAA**GGTgtgcccttgatatgctt**gttatatacaaatatccttctccagctttcatggaaagt
cttgatggtactgcatatctttaccttctgtgagctggtcctcacgtagctttctgccatggctcgact
agtgccttgctgata

F: ctgctgaacgggaaattggt (211)

R: aagcatatcaagggcacacc (203)

Amplicon: 628 bp

>DmCDC48 - Locus_2163_Transcript_2/4_Confidence_0.375_Length_351

Cell division cycle protein 48, putative, GI:110289141
agcaggatcaatgatatactggtctggttagtggcacaataataaatacagttttcttggcagacatgcc
atccatttcagtgagaagctggtttaaaacacggtccgcccaccaccagcatcaccacactgcttcc
cctctggttagcaattgaatcgagttcatcaaagaataggacacaagagctgatgacagcagccttacc
gaagatctcacgcacatttgccttcaactccccaaaccacattgtaagcaattcaggtccctttatact
aatgaagtttgctgacattcatttgcaatagccttggccaacaaagtttttccacaaccagggtgggccc
ataaaa

F: agcaggatcaatgatatactggtc

R: ctttggggccaaggctattgc

Amplicon: 323 bp

>DmATG7 - Locus_1255_Transcript_1/7_Confidence_0_682_Length_969

CGATCTAAGCAGTGGTATCAACGCAGAGTACAGGGGAAACCTAAGGTATCTCAACACCAGTAGGCAATA
ATTCAACATAGTACAGCCCCCTCTCCTTCTCCTCCCCCGGTCCGCGGTGAAATTCGCACTCTACAGTCTA
CAATCTTCTCCCTCTTCTTCTGATTGATCTCCTTGTAGACACCATGGCCAAGAGTTTCGTTCAAGC
TCCAACACCCCTTGAACACGGAGGCAGGCTGAAGCTGCACGGATCAGGGAGAAATATCCTGACAGGAT
CCCTGTTATTGTGAAAAGGCTGAAAGAAGTGACATTCGGACATTGACAAGAAGAATGACGTTGATGA
ATGTAGATATCTGGTTCCTGCTGACTTAAC**GTGGGGCAGTTTGTCTATGTG**TTAAGGAAGAGGATAAAA
GCTCAGTGCCGAGAAAGCTATATTATCTTTGTGAAGAATTATCTCCCGCCAACCGCTGCGATGATGT
CTGCTATTTACGAAGAACATAAGGATGAAGATGATTCCCTATATGAAC**TCAGTGGTGAGAACACAT**
TTGGTTCATCTTAAGCTCGAGCATTTATGCCGAGTGCATTGTGTATGTAGTAGTGTCCAAGGCAACTAC
TAATCTGGTGTATATTCTTATCATCTCCCAACTTACCGCAAGTTTCAGTACTTGGATACTCATTTCTC
CACTTTAATGTCTCTTATGGTGTGTTAATGCAACATGTATATCGTTGTATATCTCTGATTACTTCA
CTTGGTTAATCCTTTCTGACAAATTACACGAGCATTTATAAAGGAAACGATGACCAGGGAAAACCTTC
ATGCAGATTGCCAAGTACTGAAAACGAAACGAAAACAGTGAATCTGCAATAAGTTGAACCACTGAAA

ACCTTCTCATGATGGATGGCCCTCGGAGGTTTTGCAGTCGAAATTGTGACGACAATGTCGATAGTGTTG
GTTCCAGTTCTCGCGTAGCACGCCACATATTATTGTCTTTGCCTGAAGTTCAATGTATCACAAAAGCA
AACTCGTCCATTATAGAGAACTATTGAGTATGGCTGATACCATGCCCTGAACAAGGTCGGAGTCCCTGTG
ATCGGACTTGGGAGACGACGAGGAAAGGCAGAGATCCAATTTGTAGACCTTCCCATAGCTAGAAGTAT
GTATATGTTGAAGAAATTGCCACACCCCTTCCATGCCTTCTCCTCCTGCTACTGCAGTCT

F: CTTTGGTTGCATCCTGCAATAAGG
R: AGCTTCGAGAGGAAGCCAC

Amplicon: approx. 589 bp

>DmUBQ - Locus_4158_Transcript_3/4_Confidence_0.700_Length_1310 ubq
GI:102655942

AAAGGGGAGAGAAGTACTACAGCTGCAGAACATTTATGCAAACGCATTTGTATGGAGAAACCAGAAT
TGAACAAATTAAGTACAGGTAGAACACACACTTATTAATCCACAATATTTGGACCCACAATTTTCGAA
AGAGAGTCACTAAAACATGGCTAAATCACACAATATACTGATATATATCAAGAGGTAAGATGAAAGACA
TATAATAGCAAATGGTGGCTGGGCTGCAATCAAATTTCTCAGAAATCACCTCCACGAAGCGGAGGACGA
GATGAAGAGTTGATTCCTTTTGGATGTTATAGTCGGCTAGGGTTCGGCCATCTTCCAACCTGCTTCCCGG
CAAAGATGAGCCTCTGCTGGTCCGGAGGAATTCCTTCTCCTGAATTTTGGACTTCACATTATCAA
CGGTGTCCGAGCTCTCCACCTCCAAAGTGATGGTCTTCCAGTAAGAGTCTTAACAAAGATCTGCATCC
CACCACGGAGCCTCAGGACAAGGTGAAGCGTAGACTCCTTCTGAAGTTGTAATCCACGAGCGTACGGCC
ATCCTCGAGCTGTTTGCCAGCAAAGATCAGCCTCTGTTGGTCTGGAGGAATGCCCTCTTTGTCTGAT
CTTTGCCCTCACATTGTCAATTGTGTGACAGCTCTCGACCTCCAATGTGATGGTCTTGGCCGGTGAG**AGT**
CTTCACAAAGATCTGCATACCACCACGGAGACGAAGAACAAGGTGGAGTGGTATTCTTTCTGGATGTT
ATAGTCGGCAAGCGTACACCATCTTCAAGCTGTTTCCAGCAAAGATGAGCCTCTGCTGATCTGGAGGA
ATGCCCTCCTTATCCTGAATCTTGGCCTTACATTGTCAACTGTGTGACAGCTCTCCACTTCCAGGGTG
ATTGTCTT**CCAGTGAGGGTCTTGACAAAGAT**TTGCATCCCCCACGGAGGCGGAGGACAAAGGTGAAGA
GTTGATCTTTCTGGATGTTGTAGTCTGCTAGGGTTCGGCCATCTTCAAGCTGTTTCCAGCAAAGATC
AGCCTTTGCTGATCTGGGGGAATTCCTCCTTATCTTGGATCTTTGCTTTCACATTGTCAATGGTGTCA
GAGCTCTAACCTCAAGAGTGATCGTCTCCCGTTAGAGTCTTAACAAATATCTGCATCTTTTAGCAA
GAAAAGGAGAGAGAACAAGAACGGCGAGCAAACGGGGAAGCTGATCCGATCAACGACTCTGTGAAGTGT
GAATACGACCTCCGGGATGTTGAATTGATTGCCCCCCGTA**CTCTGCGTTGATACC**ACTGCTTAG

F: GTCTTCACAAAGATCTGCATACCAC
R: ATCTTTGTCAAGACCTCACTGG

Amplicon: 247

Supplementary file S2. GO terms summary. The first number is the number of genes associated to the GO term. The float in () is the percentage of the associated gene number to the total number of accumulated go counts.

Cellular Component:

17 (0.2) GO:0005576 extracellular region 2824(9.3)
 3 (0.0) GO:0044421 extracellular region part 53(0.2)
 12 (0.1) GO:0048046 apoplast 22466(73.6)
 2588 (26.3) GO:0005623 cell 22466(73.6)
 2588 (26.3) GO:0044464 cell part 0
 3 (0.0) GO:0019012 virion 0
 3 (0.0) GO:0044423 virion part 0
 72 (0.7) GO:0031974 membrane-enclosed lumen 627(2.1)
 6 (0.1) GO:0031970 organelle envelope lumen
 66 (0.7) GO:0043233 organelle lumen
 58 (0.6) GO:0031975 envelope 929(3)
 8 (0.1) GO:0030313 cell envelope
 51 (0.5) GO:0031967 organelle envelope
 746 (7.6) GO:0032991 macromolecular complex 2001(6.6)
 312 (3.2) GO:0030529 ribonucleoprotein complex
 9 (0.1) GO:0032993 protein-DNA complex 19(0.1)
 426 (4.3) GO:0043234 protein complex
 1017 (10.3) GO:0043226 organelle 17582(57.6)
 25 (0.3) GO:0031982 vesicle
 659 (6.7) GO:0043227 membrane-bounded organelle 17263
 370 (3.8) GO:0043228 non-membrane-bounded organelle
 1017 (10.3) GO:0043229 intracellular organelle
 316 (3.2) GO:0044422 organelle part 3385(11.1)
 3 (0.0) GO:0044421 extracellular region part 53(0.2)
 3 (0.0) GO:0031012 extracellular matrix
 316 (3.2) GO:0044422 organelle part 3385(11.1)
 121 (1.2) GO:0031090 organelle membrane 927
 66 (0.7) GO:0043233 organelle lumen
 316 (3.2) GO:0044446 intracellular organelle part
 3 (0.0) GO:0044423 virion part 0
 3 (0.0) GO:0019028 viral capsid 0
 2588 (26.3) GO:0044464 cell part 22466(73.6)
 1575 (16.0) GO:0005622 intracellular
 3 (0.0) GO:0008287 protein serine/threonine phosphatase complex
 2 (0.0) GO:0009349 riboflavin synthase complex
 96 (1.0) GO:0012505 endomembrane system
 1177 (12.0) GO:0016020 membrane
 1 (0.0) GO:0019008 molybdopterin synthase complex
 55 (0.6) GO:0030312 external encapsulating structure
 8 (0.1) GO:0042597 periplasmic space
 1374 (14.0) GO:0044424 intracellular part
 658 (6.7) GO:0044425 membrane part
 8 (0.1) GO:0044462 external encapsulating structure part

 Total: 43

Biological Process:

132 (1.3) GO:0044085 cellular component biogenesis 1485(4.9)
6 (0.1) GO:0032502 developmental process 4571(15.0) **underrepres**
 20 (0.2) GO:0000003 reproduction 2333(7.6)
 201 (2.0) GO:0016043 cellular component organization 3213(10.5)
 10 (0.1) GO:0016265 death 508(1.7)
 20 (0.2) GO:0022414 reproductive process 2285(7.5)
 285 (2.9) GO:0050896 response to stimulus 6503(21.3)
14 (0.1) GO:0032501 multicellular organismal process 4740(15.5)
 85 (0.9) GO:0010926 anatomical structure formation 1077(3.5)
 22 (0.2) GO:0051704 multi\organism process 1924(6.3)
733 (7.4) GO:0051234 establishment of localization 3610(11.8) **overrepresent**
 5 (0.1) GO:0022610 biological adhesion 96(0.3)
5136 (52.2) GO:0008152 metabolic process 13755(45.1)
 649 (6.6) GO:0043473 pigmentation 5455(17.9)
 753 (7.7) GO:0051179 localization 3782(12.4)
 4285 (43.6) GO:0009987 cellular process 14749(48.4)
 681 (6.9) GO:0065007 biological regulation 6426(21.1)

 Total: 17

Molecular Function:

184 (1.9) GO:0009055 electron carrier activity 525(1.7)
 86 (0.9) GO:0060089 molecular transducer activity 287(0.9)
 212 (2.2) GO:0030528 transcription regulator activity 1740(5.7)
 128 (1.3) GO:0030234 enzyme regulator activity 376(1.2)
4847 (49.3) GO:0003824 catalytic activity 8800(28.8)
 5348 (54.4) GO:0005488 binding 11327(37.1)
56 (0.6) GO:0016209 antioxidant activity 134(0.4)
 2 (0.0) GO:0016530 metalochaperone activity 5(0.0)
 14 (0.1) GO:0045735 nutrient reservoir activity 67(0.2)
 44 (0.4) GO:0045182 translation regulator activity 144(0.5)
326 (3.3) GO:0005198 structural molecule activity 550(1.8)
 423 (4.3) GO:0005215 transporter activity 1301(4.3)

 Total: 12

 Total GO terms in three ontologies: 72

